The effects of nativeness and background noise on the perceptual learning of voices and ambiguous sounds

The research reported here was supported by a Vidi-grant from the Netherlands Organization for Scientific Research (NWO; grant number: 276-89-003) awarded to Odette Scharenborg.

Published by

LOT
Trans 10
3512 JK Utrecht
The Netherlands

phone: +31 30 253 5775
e-mail: lot@uu.nl
http://www.lotschool.nl

Cover illustration by Dmitry Krizhanovsky

ISBN: 978-94-6093-267-0
NUR: 616

The effects of nativeness and background noise on the perceptual learning of voices and ambiguous sounds

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de Rector Magnificus
prof. dr. J.H.J.M. van Krieken,
volgens besluit van het College van Decanen
in het openbaar te verdedigen op
dinsdag 30 januari 2018
om 12.30 uur precies

door

Polina Aleksandrovna Drozdova

geboren 4 november 1988
te Pskov (Rusland)

Promotor:

Prof. dr. R. W. N. M. van Hout

Copromotor:

Dr. O. E. Scharenborg

Manuscriptcommisie:

Prof. dr. J. M. McQueen
Dr. S. M. Brouwer
Prof. dr. J. Vroomen (Tilburg University)
Prof. dr. S. Mattys (University of York, Verenigd Koninkrijk)
Prof. dr. F. Meunier (Université Nice Sophia Antipolis, Frankrijk)

Contents

List of Tables

List of Figures

Acknowledgements

This thesis is a result of my seven year long journey that started when I first arrived to Nijmegen to follow my master studies in 2010. A lot of people helped me on this way and without them I would have probably left the wonderful Netherlands much earlier. Throughout this journey I was lucky to work with an inspiring team of people, to whom I owe a debt of gratitude.

Odette, thank you for believing in me and offering me the position in your project four years ago! You have opened the world of psycholinguistics to me, and I really enjoyed our discussions each Wednesday morning. Thank you, also, for accepting that I am not a morning person, so that these discussions never occurred earlier than 10:00. Roeland, I am really grateful for all your advice on the statistical analysis and willingness to dig deep into my data even when I was losing hope. I have learned a lot from you, and I am still learning. Thank you for bringing structure and deadlines and putting just enough pressure at the last stage of my thesis writing. I really appreciated that both of you were always responsive to my questions and e-mails, whether on weekdays, weekends or evenings.

During my project, I was lucky to have an opportunity to run my experiments at the university of York in the laboratory of Prof. dr. Sven Mattys. Sven, thank you and your team (especially Dorina and Huarda) for allowing me to use your facilities and for your help in all practical arrangements, and thank you for agreeing to be a member of my thesis committee.

My work environment would not have been this great without my colleagues on the 8th floor. Helmer and Catia, thank you for your support during my research master and especially during the year between

my research master and the start of my PhD: without your help and reassurance I would have given up already after the first failed proposal. Juul, Jule, Florian and Themis, I really enjoyed our project dinners and lunches and our exchange of ideas. It was great to have people around to share both good and difficult times!

I am grateful to my colleagues for reminding me that a PhD project is not only about writing and thinking while staring at the computer screen, but also about fun! I will try to mention all of you, whether you are still on the 8th floor or not! Bart, Claire, Erwin, Ferdy, Henk, Job, Louis, Mario, Marjoke, Michele, Micha, Remy, Steve, Thijs, Vanja, Wessel, Xiaoru and many others: it was great to share coffee breaks and lunches with you and also meet for fun outside the office for Christmas dinners, board games, movies and festivals. I am also happy to have the best possible paranymphs: Eric and Sara. Thank you for sharing this special day with me! Mario, I would also like to thank your sister, Natalie, for translating the summary of my dissertation into Dutch.

I have managed to run a lot of experiments during my PhD. A special thank you goes to Joop for creating all these handy PRAAT scripts which I have extensively used in my project and helping me in performing multiple phonetic analyses. I am grateful to the CLS lab team, especially to Margret van Beuningen for arranging CLS lab lunches, lab bookings and IRIS checks. Thanks to you, my testing always went smoothly. All our student assistants require an additional mentioning here! Esther, Jiska, Marloes, Margot, Tijn and Vincent, thank you for testing so many people for me. Without you, finishing my experiments would have taken me much longer. I have to mention, Alastair Smith here, who has officially become the voice of our project. I really appreciate that you were always ready to come to our rescue when we needed to record a native speaker of English.

It was very helpful to discuss my experimental design and results, and practice for conference presentations during PI-group and Sound Learning meetings. Dear fellow researchers, your input and comments helped to see my research from a different angle and thoroughly plan my experiments. I would also like to thank IMPRS, and especially Kevin, Dirkje and Els for organizing useful courses and fun activities for PhD students. It was always interesting to meet other people from IMPRS cohorts and share our experiences.

During my student life in Nijmegen, before starting my PhD project, I was lucky to meet different wonderful people who made my social life

outside the University complete, and who were always there for me and were always ready to listen to me and support me throughout my Dutch life and PhD endeavors! Fatima, Gaby, Ingrid, Nastya, Sophie, Uli and Veronica, thank you for all the moments we shared together! I am pretty sure that there are many more to come no matter the distance. Finally. I have a group of people back in my hometown, Pskov, who were always happy for my achievements and supported me in everything even though there are thousands of kilometers between us. Валя, Катя, Ксюша, Настя Д. и Настя Ц., спасибо вам за вашу поддержку на протяжении уже стольких лет!

And I definitely would not have been where I am now, if it were not for my parents! Мама и папа, спасибо, что поддержали моё решение уехать учиться за границу, что всегда поддерживаете и принимаете путь, который я выбираю. Спасибо, что гордитесь мной!

The last, but not least, I want to thank my friend, my partner, and my fiancé, Ilya, who had to cope with all my PhD ups and downs, practice my presentations with me and check my articles (although always complaining about their length). Thank you for your love and support and especially for always pushing me to achieve new heights and learn new things.

CHAPTER 1

Perceptual learning in adverse conditions

1.1 Introduction

No two realizations of the same word or even sound are identical. While listening to speech, listeners thus have to deal with a highly variable speech signal. This variability is caused by many different factors, such as speaker-related idiosyncrasies, speaking style, dialects, emotional states, age, gender and size of the vocal tract of the speaker, in other words, pronunciation variation within and between speakers. Despite this variability, native listeners are usually able to correctly interpret the speech signal. How do listeners arrive at the correct interpretation? Existing theories of speech comprehension postulate that listeners map the variable signal onto pre-existing representations (Marslen-Wilson & Tyler, 1980; McQueen, 2005; Weber & Scharenborg, 2012). Mapping this highly variable signal onto pre-existing representations is easier when the listener is tuned into the speech of the speaker. This adaptation to unfamiliar speakers' pronunciations, a form of perceptual learning (Norris, McQueen, & Cutler, 2003), improves speech perception (Goh, 2005) and word recognition (Nygaard & Pisoni, 1998).

Perceptual learning was first defined by Gibson (1963, p.29) as a relatively permanent and consistent change in the perception of stimuli following practice or experience with those stimuli, and has been demon-

strated in different modalities: visual, olfactory, tactile, and auditory (Samuel & Kraljic, 2009). This thesis is concerned with perceptual learning in the auditory modality and focuses on two lines of research each falling into one of the two themes of perceptual learning defined by Samuel and Kraljic (2009). One line of research focuses on retuning of phonetic categories as a result of exposure to consistent ambiguous input in the form of visual (Bertelson, Vroomen, & De Gelder, 2003; Van Linden & Vroomen, 2007) or lexical information (Norris et al., 2003). In this thesis we investigate retuning of phonetic category boundaries evoked by lexical knowledge, termed lexically-guided perceptual learning by Norris et al. (2003). The other line of research focuses on the improvement in listeners' ability to comprehend speech as a result of exposure to and familiarization with previously unfamiliar talkers (see Samuel & Kraljic, 2009, for a review).

In everyday listening, speech comprehension often occurs in listening conditions that are not optimal. This thesis focuses on two types of non-optimal listening conditions: the presence of background noise and imperfect language knowledge due to listening in a non-native language (Garcia Lecumberri, Cooke, & Cutler, 2010; Mattys, Davis, Bradlow, & Scott, 2012). Both non-nativeness and the presence of background noise result in additional processing cost in speech comprehension, slowing down word recognition and decreasing recognition accuracy (Brouwer & Bradlow, 2011; Weber & Cutler, 2004). Being able to tune into a speaker might then be beneficial for speech comprehension, especially when listening conditions are really bad (McLennan & Luce, 2005; Nygaard & Pisoni, 1998). However, masking of the speech signal due to the presence of noise or listener's reduced knowledge of the non-native language might interfere with perceptual learning, as listeners might have difficulties picking up acoustic, phonological and lexical information from the signal important for perceptual learning (Norris et al., 2003; Perrachione, Del Tufo, & Gabrieli, 2011; Schacter & Church, 1992).

The aim of the present thesis is to study the role of nativeness and the presence of noise in dealing with the variation in the speech signal introduced by variability within and between speakers. Chapters 2 to 4 present experiments devoted to the adaptation of listeners' perceptual system to talker's ambiguous pronunciations in native and non-native listening in clean and in the presence of background noise by means of the lexically-guided perceptual learning paradigm. Chapters 5 and 6 are devoted to the adaptation of the listeners' perceptual system to talkers'

voices and the potential benefit of this adaptation on both speech processing and word recognition in clean and in the presence of background noise in non-native listening. The remaining part of this introduction is as follows. Section 1.2 discusses spoken word recognition in general and the difficulties associated with speech recognition in the presence of background noise and in a non-native language. Section 1.3 reviews the existing literature on lexically-guided perceptual learning. Section 1.4 reviews the studies on the adaptation to talkers' voices and the talker familiarity benefit. Section 1.5 explains the research questions that are central to this thesis. Finally, the general methodology and the outline of the present thesis are described in Sections 1.6 and 1.7, respectively.

1.2 Spoken word recognition

The speech signal combines both linguistic information of what is being said and indexical information of who said it (Abercombie, 1967). At the heart of the spoken word recognition process is the mapping of the variable speech signal onto pre-existing discrete representations (Marslen-Wilson & Tyler, 1980; McQueen, 2005; Weber & Scharenborg, 2012). The abstract or more detailed nature of these representations, the levels of processing and the flow of information between these levels are however still under debate.

Abstractionist theories of speech comprehension postulate that the speech signal is mapped onto abstract representations at the prelexical level of processing and subsequently onto lexical representations at the lexical level (Marslen-Wilson & Tyler, 1980; McClelland & Elman, 1986; McQueen, Cutler, & Norris, 2006; Norris, 1994). These theories hypothesize that indexical information is not stored in the mental lexicon of the listener but rather is abstracted away at the prelexical level of processing (Cutler, 2010; Cutler, Eisner, McQueen, & Norris, 2010).

Episodic theories (Goldinger, 1996, 1998) assume the existence of a vast collection of memory traces for each known word. According to the episodic theory of speech perception, each perceptual episode is encoded in the memory of the listener including much of its phonetic detail (e.g., talker-specific indexical information). Upon hearing a new speech input, this speech input is directly compared to all stored exemplars, thus without the need for a prelexical abstraction before accessing the lexicon.

Despite their differences, all theories of human speech processing

assume that upon hearing a speech signal, multiple lexical candidates (abstractionist theories) or traces (episodic theories) are activated, and compete for recognition. Simply put, the winning candidate or trace is the one matching the speech input best (see Weber and Scharenborg, 2012 for an overview of existing computational models of spoken word recognition).

Listening in a non-native language is typically harder than listening in a native language. There are a number of reasons for this difference. Firstly, phonetic categories and contrasts present in the non-native language might be absent or realized differently from those in the native language of the listener (Flege, 1995). Due to the subsequent imperfect sound perception, more words in the non-native language compete for recognition (Broersma, 2012), in addition to spuriously activated words from the native language of the listener (Weber & Cutler, 2004). This increased competition slows down word recognition and decreases word recognition accuracy (Norris, McQueen, & Cutler, 1995). Secondly, lexical knowledge in the non-native language is impoverished in comparison to native lexical knowledge (Garcia Lecumberri et al., 2010). The word in the non-native language might even be missing from the lexical repertoire of the non-native listener, and, if so, word recognition is severely hampered.

Multiple studies have shown that the presence of background noise in the speech signal results in an additional processing cost compared to listening in quiet listening conditions, even in native listening (e.g., Brouwer & Bradlow, 2011, 2016), which manifests itself as a slowing down of the speech recognition process and an increase in recognition errors. This is due to several reasons. For one, listeners have to segregate the speech signal from the noise signal (Brouwer & Bradlow, 2016). Second, in the presence of noise more words compete for activation than in clean listening conditions (Brouwer & Bradlow, 2011; Scharenborg, Coumans, & van Hout, 2017). Moreover, the presence of intermittent noise in the speech signal makes listeners relatively less certain about what words they have heard (McQueen & Huettig, 2012).

When listening in noise in a non-native language, listeners have to deal with both an imperfect signal and imperfect lexical and phonological knowledge (Garcia Lecumberri et al., 2010). While native listeners can rely on sentence context and prosody (if available), non-native listeners have been shown to use contextual information to a lesser extent when the signal is noisy than native listeners do (e.g., Mayo, Florentine, &

Buus, 1997; Meador, Flege, & Mackay, 2000; Scharenborg, Kolkman, Kakouros, & Post, 2016). Moreover, there is evidence that noise impacts the processes underlying spoken word recognition in non-native listeners more than in native listeners (e.g., Scharenborg et al., 2017).

1.3 Adaptation to a talker's pronunciation

Lexically-guided perceptual learning was first demonstrated in the seminal study by Norris and colleagues (Norris et al., 2003). Native Dutch listeners in that study were exposed to an ambiguous sound halfway between /f/ and /s/. For one group of listeners this sound substituted /s/ in an exposure phase, while all words containing /f/ were natural. For another group all /f/ sounds were substituted by the ambiguous sound, while all words with /s/ remained natural. In a subsequent task, listeners had to categorize stimuli on an /f/-/s/ continuum as either containing /f/ or /s/. The group exposed to the ambiguous sounds in the words with /s/ categorized significantly more ambiguous stimuli as /s/ than the other group. This difference between the two groups of listeners is termed the lexically-guided perceptual learning effect and can be explained by a temporal adjustment of the phonetic category boundaries (Clarke-Davidson, Luce, & Sawusch, 2008).

This adjustment is guided by lexical knowledge (Norris et al., 2003) and knowledge about the phonotactics of the language (Cutler, McQueen, Butterfield, & Norris, 2008) of the listeners, which is shown by the fact that learning happens only when ambiguous sounds are embedded in real words or are part of phonotactically legal sequences of the languages. The importance of lexical information for the phonetic category retuning to occur was also underlined by a study by Jesse and McQueen (2011) who found that retuning only occurs when enough lexical information is available: no lexically-guided perceptual learning occurred when ambiguous sounds were located at the beginning of the words. Moreover, Scharenborg and Janse (2013) found that listeners who accepted more words containing the ambiguous sound as real words during a lexical decision task in the exposure phase showed more category retuning in the subsequent phonetic categorization task. Note, however, that Chládková, Podlipskỳ, and Chionidou (2017) demonstrated perceptual adjustment of phonetic category boundaries for Greek vowels /i/-/e/ and /u/-/o/ irrespective of whether these vowels were embedded in real words or in

non-words, suggesting that at least for vowels (in contrast to the earlier studies on consonant retuning) phonetic retuning is possible outside a lexical context.

Lexically-guided perceptual learning has been demonstrated in native listening with different sound contrasts: fricatives: /s/ and /f/ (e.g., Norris et al., 2003; Eisner & McQueen, 2006), /s/ and /ʃ/ (Kraljic & Samuel, 2005); stops: /d/ and /t/ (Kraljic & Samuel, 2006), /p/ and /t/ (Van Linden & Vroomen, 2007), liquids : /l/ and /r/ (Scharenborg, Mitterer, & McQueen, 2011), and tones (Mitterer, Chen, & Zhou, 2011), and not only in adults, but also in children (McQueen, Tyler, & Cutler, 2012) and older listeners (Scharenborg & Janse, 2013). Learning was observed when listeners had to pay attention to the lexical status of the words during exposure (e.g., in a lexical decision task) but also when they were passively listening to a short story (Eisner & McQueen, 2006) or simply counting the number of words (McQueen, Norris, & Cutler, 2006).

Although lexically-guided perceptual learning occurs relatively fast (as few as ten exposure items are enough for the retuning to happen (Kraljic & Samuel, 2007; Poellmann, McQueen, & Mitterer, 2011), items with ambiguous sounds are not immediately perceived as real words. A number of studies (Scharenborg & Janse, 2013; Schuhmann, 2016) demonstrated that when a lexical decision task is used in the exposure phase, items with ambiguous sounds are judged less often as real words than their counterparts with the same sounds in their natural form. However, the acceptance rates were shown to increase over the course of the exposure phase (Scharenborg & Janse, 2013) indicating integration of the ambiguous sound in an existing sound category. The time-course of lexically-guided perceptual learning is the topic of investigation in Chapter 4 of the present thesis.

Lexically-guided retuning is hypothesized to be beneficial for listeners as it facilitates the recognition of future speech input containing the same ambiguous sound patterns (McQueen, Cutler, & Norris, 2006). This generalization to words not presented to the listeners during exposure suggests the need for abstract representations at a prelexical processing level (McQueen, Cutler, & Norris, 2006), and as such is taken as evidence against episodic theories of spoken word recognition (McQueen, Cutler, & Norris, 2006).

1.3.1 Lexically-guided perceptual learning in non-native listening

Given the role of lexical and phonotactic information in lexically-guided perceptual learning, at least for consonants, it may be presumed that lexical and phonological knowledge of the listeners and the ability to retrieve this information from the signal are prerequisites for retuning to occur. This might be problematic for non-native listeners whose lexical representations might be impaired (Garcia Lecumberri et al., 2010). Moreover, as pointed out before, non-native listeners might not have phonetic categories of sounds in the non-native language that are specific enough to allow retuning to occur.

Only a limited number of studies addressed the question of lexically-guided perceptual learning by non-native listeners. Reinisch, Weber, and Mitterer (2013), for instance, demonstrated retuning of /s/-/f/ by German listeners who were highly proficient non-native listeners of Dutch when listening to Dutch. However, as argued by Reinisch et al. (2013), it was not clear whether the German listeners retuned their native German or non-native Dutch sound categories, since /s/ and /f/ are highly similar in Dutch and German. Schuhmann (2016) showed that native English and native German listeners with knowledge of German and English, respectively, demonstrated lexically-guided perceptual learning in both their native and non-native language when the /s/-/f/ contrast, phonetically highly similar in English and German, was used, thus showing cross-language lexically-guided perceptual learning. These results indeed seem to suggest that highly phonetically similar sounds in a non-native language can retune (at least) native sound categories. Hanulikova and Ekstörm (2017) investigated lexical adaptation by native German listeners and Swedish and Finnish non-native listeners of German to a novel accent in German, which was created by lowering front and back vowels. Both Swedish and Finnish listeners were proficient in German. Acceptance scores for ambiguous words containing the lowered vowels were compared before and after an exposure story. Only German and Swedish listeners demonstrated a significant increase in acceptance scores, which was explained by the authors by the lexical and phonological similarities between German and Swedish.

These studies show that ambiguous sounds in a non-native language can cause the adaptation of native phonetic categories. Chapter 2 of the present thesis investigates whether ambiguous sounds in a non-native language can lead to the adaptation of *non-native* phonetic category boundaries.

1.3.2 Lexically-guided perceptual learning in noise

The ability to quickly adapt phonetic category boundaries as a result of exposure to ambiguous input shows the flexibility of the human perceptual system (Cutler, 2012). However, there are clear bounds to this flexibility. No retuning occurs when the ambiguity is characteristic for a certain dialect (Kraljic, Brennan, & Samuel, 2008) or when the speaker has a pen in his/her mouth (Kraljic, Samuel, & Brennan, 2008). Samuel and colleagues (Samuel, 2011; Samuel & Kraljic, 2009; Zhang & Samuel, 2014) hypothesized that the perceptual system only adapts to reliable changes that are likely to persist and that retuning is blocked when listeners cannot attribute the ambiguity in the speech signal to idiosyncrasies in the production of a particular speaker. This principle for relative stability and flexibility of the human perceptual system depending on the context in which new information occurs was coined the Conservative Adjustment/Restructuring Principle (Samuel & Kraljic, 2009).

Background noise increases the ambiguity of the speech signal. On the one hand, it has been shown to interfere with the competition process. McQueen and Huettig (2012) found that listeners became less certain about what words they heard when the speech signal contained stretches of noise. The listeners in their eye-tracking study changed their eye fixation behavior (looked less at onset-overlapping words compared to a clean condition) when the target word was placed in a sentence with intermittent noise compared to a sentence in the clean. On the other hand, the presence of background noise can change the acoustics of the sound through masking of the acoustic cues in the speech signal or through perceptually attaching spectro-temporal information to the sound (e.g., Cooke, 2009), which might interfere with the retrieval of (correct) lexical and phonological information from the speech signal. Only one study so far investigated the effect of noise on lexically-guided perceptual learning. Zhang and Samuel (2014) showed that lexically-guided perceptual learning in native listening is blocked when the stimuli containing the ambiguous sounds are embedded in background noise, even though the

noise was not placed on the actual critical ambiguous sounds and the presence of noise did not prevent listeners from correctly interpreting and transcribing the stimuli. The authors suggested that the presence of background noise increased the overall variability in the speech signal because of which the variability of the ambiguous sound could no longer be interpreted as a reliable cue to trigger retuning. Chapter 3 of this thesis will further investigate the effect of background noise on the retuning of phonetic categories.

1.4 Adaptation to a talker's voice

In everyday life, people are exposed to multiple talkers. The voices of talkers differ in multiple voice-quality features such as, e.g., loudness, pitch, phonation types and nasalization (Laver, 1968). These features define various talker-related (indexical) information such as gender, age, size and emotional state. Abstract and episodic theories of speech perception hypothesize that this talker-related information is either processed separately from (abstract theory) or together with (episodic theory) the linguistic information in the speech signal, with past research providing evidence for both separate and interacting processing (Winters, Levi, & Pisoni, 2008). In the present thesis, the interaction of language-specific and language independent information in the signal is investigated by studying several factors influencing voice learning, i.e., learning to recognize previously unfamiliar voices (Chapter 5; Perrachione & Wong, 2007; Winters et al., 2008; Zarate, Tian, Woods, & Poeppel, 2015) and studying whether familiarity with the voice of the talker influences speech processing and word recognition (Chapter 6; Goh, 2005; Nygaard & Pisoni, 1998; Palmeri, Goldinger, & Pisoni, 1993), focusing specifically on non-native listeners.

1.4.1 Language-specific and talker-specific factors influencing voice learning

Learning to differentiate and recognize a certain voice entails learning to associate acoustic properties with this particular voice. Different acoustic characteristics of voices were shown to play a role in voice familiarization, such as pitch, formant frequencies, jitter and shimmer, with certain characteristics being more important depending on the gender of the speaker and proficiency of the listener (expert on non-expert) (see Baumann &

Belin, 2010 for an overview). Moreover, acoustic characteristics of talkers' voices were shown to relate to how easy or difficult those voices are to memorize and differentiate, i.e., more distinct voices are easier to remember (Levi, 2014; Papcun, Kreiman, & Davis, 1989). The importance of these acoustic cues is further shown by voice learning studies demonstrating that listeners can identify talkers in the absence of intelligible linguistic content as in time-reversed speech (Bricker & Pruzansky, 1966; Sheffert, Pisoni, Fellowes, & Remez, 2002) and in a completely unfamiliar language (Winters et al., 2008).

The finding that voice recognition not only depends on acoustic, language-independent characteristics of the voice was first presented in the study by Goggin and colleagues (Goggin, Thompson, Strube, & Simental, 1991). Monolingual English listeners in that study identified bilingual English-German speakers better when these speakers spoke English than when they spoke German, while the reverse pattern was true for monolingual German listeners. This native language advantage was later confirmed using voice-learning experiments which showed that listeners learned voices better in their native language than in a non-native (but familiar) language, while the worst performance was obtained for listeners learning to recognize voices in an unfamiliar language (Bregman & Creel, 2014; Perrachione & Wong, 2007).

The nature of this native language advantage was suggested to be related to listeners' phonological knowledge (Perrachione et al., 2011; Zarate et al., 2015). Perrachione et al. (2011) and Jimenez (2012) showed that voice learning and discrimination performance is correlated with the results of a test on phonological processing in both dyslexic and healthy populations. Zarate et al. (2015) further demonstrated that voice recognition performance of the listeners significantly improves when additional phonological information becomes available. In their study, native English listeners were trained to recognize several talkers on the basis of several types of stimuli: non-speech vocal sounds (e.g., laugh, cry), words in Mandarin, German or English, or English non-words composed of syllables from existing English words. Voice recognition after training on non-speech was the worst, but still above chance, followed by performance after training in Mandarin, an unfamiliar and typologically distant language compared to the participants' native language. Interestingly, voice recognition performance did not differ across condition blocks with high phonological familiarity (German, pseudo-English, and English), despite the fact that German was unfamiliar to the listen-

ers. The authors concluded that it is not a general language familiarity, but rather phonological familiarity that plays a role in voice recognition, in addition to acoustic characteristics of the voice, since listeners were able to learn to recognize voices on the basis of non-speech.

Summarizing, previous studies have shown that learning to recognize a voice depends on a number of factors, such as acoustic characteristics of the voices, phonological information in the signal, and listeners' ability to retrieve and use this information during voice learning. Factors influencing voice recognition in non-native listening are investigated and discussed in Chapter 5 of the present thesis.

1.4.2 Same talker and familiar talker benefit in speech processing

Voice-specific information has been shown to be used by listeners to their advantage in speech comprehension. Palmeri et al. (1993) showed that repeated words are recognized faster and more accurately when these words were produced by the same talker (also referred to as a same-voice repetition) than when they were produced by a different talker (different-voice repetition). Goldinger (1996) observed the effect of same-voice repetition in both a recognition memory task (similar to Palmeri et al., 1993) and a word identification task, where listeners had to identify words in white noise. These observations gave rise to episodic theories of lexical access (Goldinger, 1998), postulating that for every known word a vast collection of traces exists in the memory of the listeners (see also Section 2). These traces are presumed to contain detailed information from the speech signal including indexical information. Upon hearing a new word, all stored traces are activated depending on their similarity with the current word. The same talker advantage is then caused by a full match in surface detail between the just heard word and the stored trace.

Goh (2005) showed that speech processing can be facilitated not only when there is a full match (same talker, same word), but also when the voice of the talker is familiar (same talker, different word). Listeners in that study were more accurate identifying words as already presented when these words were produced by familiar talkers (who listeners had already heard during the exposure phase of the experiment) than when these words were produced by new talkers. Nygaard, Sidaras, and Alexander (2008) furthermore showed that listeners are faster at repeat-

ing words when these words were produced by familiar than unfamiliar talkers. This familiar talker benefit was also demonstrated using a word recognition task when explicit training on the voices of the talkers was provided. Nygaard, Sommers, and Pisoni (1994) showed that listeners who had been trained to recognize the voices of previously unfamiliar talkers over a period of nine days outperformed an untrained group of listeners in a subsequent word recognition task with the words embedded in noise at different signal-to-noise ratios (see also Nygaard & Pisoni, 1998; Yonan & Sommers, 2000).

Clearly, indexical information is not (fully) removed from the speech signal during listening but rather is stored in the memory of the listeners and can facilitate both speech processing and word recognition.

1.4.3 Same talker and familiar talker benefit in non-native listening

Given that listening in a non-native language is harder than listening in a native language, non-native listeners could potentially benefit from using information about the talker's voice to improve speech comprehension in the non-native language. Although non-native listeners were shown to be able to learn the voices speaking in a non-native language, there is only a limited set of studies investigating whether non-native listeners are able to use indexical information during speech processing and word recognition. Same talker benefit in non-native speech processing was demonstrated by Trofimovich (2005) and Winters, Lichtman, and Weber (2013). Trofimovich (2005) showed that learners of Spanish respond faster to the repeated Spanish words than the new words only when these words are produced by the same talker as in exposure. Furthermore, Winters et al. (2013) demonstrated that English learners of German recognize that the word was already presented more accurately when this word is repeated in the same voice than when this word is repeated in a different voice. Non-native listeners with knowledge of the language, like native listeners, thus seem to have a same talker benefit in speech processing.

No studies to our knowledge specifically look at the familiar talker benefit in non-native listening. Levi and colleagues (Levi, Winters, & Pisoni, 2011) showed that listeners are able to learn to recognize voices in an unfamiliar language but they did not observe a familiar talker benefit during word recognition with these voices. They suggested that

the familiar talker benefit depends on the language knowledge of the listeners and the knowledge of how a talker produces linguistically relevant contrasts in a particular language. Non-native listeners with knowledge of the non-native language are expected to be able to establish acoustic-phonetic links between talker information and what is being said during the training, which could lead to the emergence of a familiar talker benefit. The role of familiarity with the voice of the talker on non-native speech processing and word recognition is further investigated in Chapter 6 of the present thesis.

1.4.4 Same talker and familiar talker benefit in noisy listening conditions

Noise has typically been assumed to increase perceptual difficulty, slowing down speech comprehension and requiring increased attention from the listener. The time-course hypothesis, introduced by Luce, McLennan, and Chance-Luce (2003) and its attention-based extension introduced by Maibauer, Markis, Newell, and McLennan (2014) posit that indexical effects appear to emerge relatively late in processing, unless the listeners pay increased attention to the stimuli due to, e.g., familiarity with the voice (Maibauer et al., 2014) or explicit focus on the voice of the speaker imposed by the task itself (Theodore, Blumstein, & Luthra, 2015). Both speed of processing and attention modulate the emergence of the effects of indexical information according to these hypotheses (Tuft, McLennan, & Krestar, 2016). If these hypotheses are true then facilitation from the same or a familiar talker's voice should be observed to a greater extent in noise than in clean.

The results of different studies on the effect of the presence of noise on the same talker benefit are however not clear-cut. Schacter and Church (1992), for instance, did not find a same talker benefit when words were presented in clean listening conditions in the exposure phase and subsequently had to be identified in noise in the test phase. On the other hand, (Goldinger, 1996) observed facilitatory effects of same voice repetitions in a word recognition task when both study and test phases were in noise, which made the author suggest that noise can be encoded alongside other surface details of the stimuli (including voice). Moreover, Nijveld, ten Bosch, and Ernestus (2015) found that listeners reacted faster to items produced by the same talkers as in exposure than to items produced by different talkers only when the stimuli were embedded in noise

compared to when they were presented in clean, even though listeners were in general faster in the noise listening condition than in the clean listening condition. The familiar talker benefit was also observed when listeners had to identify words in noise (Nygaard et al., 1994; Nygaard & Pisoni, 1998), with a larger benefit from familiarity with a talker at increasingly difficult single-to-noise ratios (Nygaard & Pisoni, 1998; Yonan & Sommers, 2000). It is therefore not entirely clear what role background noise plays on the use of indexical information during non-native speech comprehension. This question is further investigated in Chapter 6 of the present thesis.

1.5 Research questions

Previous studies thus showed an important role for perceptual learning in dealing with the variation in the speech signal caused by variability within one and between speakers and the presence of noise. Perceptual learning studies devoted to lexical retuning of phonetic boundaries and the learning of talkers' voices and its effect on speech processing demonstrated that neither extreme abstract nor episodic theories of lexical access can account for the available experimental data. Episodic theories of lexical access cannot account for the generalization of the lexically-guided perceptual learning effect and extreme abstract theories cannot explain the interaction of indexical and linguistic information during speech comprehension. As stated by McLennan (2007), the use of abstract and/or episodic representations is likely to vary depending on listening situations and the type of listener. Therefore, studying the combined effect of noise and non-nativeness may give more insight into the interaction of linguistic and indexical information in human processing of the variable speech signal.

The overall aim of the present thesis is to study the role of nativeness and the presence of noise in dealing with the variation in the speech signal introduced by variability within and between speakers. We focus on perceptual learning as a means of dealing with variability in the speech signal. More specifically, this thesis focuses on lexically-guided perceptual learning, used by listeners to adapt to ambiguous pronunciations of a particular speaker, and perceptual learning of voices, which is hypothesized to facilitate both speech perception and word recognition. The experiments described in this thesis are conducted not only in optimal

but also in adverse listening conditions, i.e., when the speech signal is noisy and/or the listener has imperfect lexical and phonological knowledge of a non-native language. This leads to the following three research questions that will be addressed in this thesis:

1. How does lexically-guided perceptual learning function in native and non-native listening in both clean (Chapters 2, 4) and noisy listening conditions (Chapter 3)?

2. What factors influence perceptual learning of voices in non-native listening (Chapter 5)?

3. Does perceptual learning of voices facilitate speech comprehension in non-native listening in both clean and noisy listening conditions (Chapter 6)?

1.6 Methodology

The experiments in this thesis used a variety of behavioral perceptual tasks conducted with Dutch non-native listeners of English, native Dutch listeners, and native British listeners. All experiments contain an exposure and a test phase. For the lexically-guided perceptual learning experiments presented in Chapters 2 to 4, participants were exposed to an ambiguous pronunciation of a certain sound from one speaker either in the form of a lexical decision task (similar to Norris et al., 2003; Chapter 4) or in the form of passive listening to a short story (similar to Eisner & McQueen, 2006; Chapters 2 and 3). The exposure phase was conducted in noise-free listening conditions (Chapters 2 to 4) or with added intermittent background noise (Chapter 3). After the exposure phase, a phonetic categorization task in clean listening conditions was used in the test phase to investigate whether participants retuned their phonetic boundaries.

Chapters 5 and 6 describe a voice learning study in which participants were familiarized with the voices of four talkers over the course of four days (similar to Nygaard & Pisoni, 1998; Chapter 5). To study whether familiarity with the voice of the talker improved speech processing and word recognition, recognition memory tasks and word recognition tasks were carried out by the same participants who participated in the voice learning experiment (Chapter 6).

All experiments described in this thesis included a measure of proficiency in the non-native language which was obtained using an unspeeded lexical decision task in English, LexTALE (Lemhöfer & Broersma, 2012). The study described in Chapter 5 additionally includes measures of phonological aptitude and working memory for each individual listener.

1.7 Outline

Chapter 2 addresses the question **whether an ambiguous sound in a non-native language can retune non-native phonetic categories**. This question was investigated using the British English sound contrast /l/-/ɹ/. The /l/-/ɹ/contrast, where the /ɹ/ is realized differently in Dutch and English, allowed us to test the hypothesis that non-native phone categories can be adapted on the basis of non-native ambiguous speech. This is in contrast to earlier studies on lexically-guided perceptual learning in non-native listening which focused on phonetic categories which are also present in the native language of the listeners. Three groups of listeners were tested: British English listeners, Dutch non-native listeners of English, and a control group of Dutch listeners.

Chapter 3 investigates the **effect of intermittent noise on lexically-guided perceptual learning in both native and non-native listening**. The same experimental set-up was used as in Chapter 2 but this time stretches of noise were added to the short story. These stretches of noise were never placed on the words containing the ambiguous sounds. It was hypothesized that intermittent noise in the short story would impede lexically-guided perceptual learning, and more so in non-native compared to native listening. This hypothesis was investigated in four experiments with native English and Dutch non-native listeners of English in two listening conditions, i.e., in the clean and in the presence of background noise.

Chapter 4 investigates **in how far items containing ambiguous sounds are perceived and processed as real words**. Since perceptual learning implies inclusion of an ambiguous sound in an existing phonetic category, it was expected that processing and recognition of the words containing the ambiguous sounds would become more word-like with increasing exposure to the ambiguous sound. To that end, the time-course of accepting words containing an ambiguous sound as a word and the spreading of activation to semantically-related words of these

words containing ambiguous sounds were investigated. The exposure phase of the lexically-guided perceptual learning experiment consisted of an auditory semantic priming task embedded in a standard lexical decision task. The experiment in this chapter was conducted with Dutch native listeners in Dutch in clean listening conditions.

Chapter 5 investigates **to what extent different speaker-, stimuli- and listener-related factors influence voice learning and voice recognition in non-native listening**. To that end, Dutch non-native listeners of English learned to recognize the voices of four native English speakers speaking in English during a four-day training period. The contribution of speaker-related characteristics, such as fundamental frequency and average word length, stimuli-related characteristics, such as sound composition and lexical frequency of words, and listener-related characteristics, such as lexical knowledge, phonological aptitude and working memory to voice recognition accuracy were investigated. Voice learning in this chapter was conducted in the clean.

Chapter 6 investigates **the role of familiarity with the voice of the talker in non-native speech processing and word recognition in both clean and noisy listening conditions**. Preceding and following the voice training task described in Chapter 5, all listeners performed two additional tasks: a recognition memory task and a word recognition task. There were some technical problems with these tasks with 10 participants. These participants were not included in the analyses of Chapter 6. The role of voice familiarity in speech processing was studied by means of an explicit recognition memory task on each day of the voice learning experiment described in Chapter 5. In this task, participants had to indicate whether they already heard the presented word (presented either in the clean or embedded in background noise) in the voice learning task (on that same day). The performance of the listeners on the words produced by familiar talkers (i.e., the talkers they were trained on) and by new, unfamiliar talkers were compared. The effect of familiarity with the voice of the talker on word recognition was studied in a word recognition task with various levels of noise with one group of the listeners performing the task with the voice of the familiar talker, while the other group was performing the task with the voice of an unfamiliar talker. We hypothesized the emergence of a talker familiarity benefit in both non-native speech processing and word recognition with a larger benefit for adverse listening conditions with higher levels of background noise.

Chapter 7 discusses the findings from previous chapters in relation to the three research questions. Conclusions are formulated on the basis of these findings and implications for existing theories of speech perception and possible lines for future research are discussed.

CHAPTER 2

Lexically-guided perceptual learning in non-native listening

This Chapter has been adapted from
Drozdova, P., van Hout, R., & Scharenborg, O. (2016). Lexically-guided perceptual learning in non-native listening. *Bilingualism: Language and Cognition*, 19(5), 914-920.

Abstract

There is ample evidence that native and non-native listeners use lexical knowledge to retune their native phonetic categories following ambiguous pronunciations. The present study investigates whether a non-native ambiguous sound can retune non-native phonetic categories. After a brief exposure to an ambiguous British English [l/ɹ] sound, Dutch listeners demonstrated retuning. This retuning was, however, asymmetrical: the non-native listeners seemed to show (more) retuning of the /ɹ/ category than of the /l/ category, suggesting that non-native listeners can retune non-native phonetic categories. This asymmetry is argued to be related to the large phonetic variability of /r/ in both Dutch and English.

2.1 Introduction

The speech signal is variable: speakers pronounce sounds differently depending on, e.g., their gender, dialect, accent, and age. Listeners cope with this variation by quickly tuning into a speaker, even when pronunciations are ambiguous (Norris et al., 2003). There is ample evidence that native speakers use lexical and phonotactic knowledge to retune their phonetic categories in response to ambiguous pronunciations of sounds (see Samuel & Kraljic, 2009, for an overview), and apply this learning to novel items (McQueen, Cutler, & Norris, 2006). This competence, termed lexically-guided perceptual learning (Norris et al., 2003), leads to temporary adjustments of phonetic category boundaries (Clarke-Davidson et al., 2008). We argue here that highly proficient non-native listeners are able to benefit from the same process and, as a result of a brief exposure to an ambiguous sound, can retune their second language (L2) phonetic category boundaries to include this ambiguous sound, and can apply this learning to new, unheard words.

Lexically-guided retuning has been demonstrated for native listeners using an exposure-test paradigm. In the seminal study by Norris et al. (2003), Dutch listeners were exposed to word items containing an ambiguous sound between /f/ and /s/. The authors demonstrated that listeners exposed to the ambiguous sound in /f/-final words (e.g., gira[f/s], where *giraffe* is an existing Dutch word and *giras* is not) learned to interpret the sound as /f/. The group exposed to the ambiguous sound in /s/-final words (e.g., mui[f/s], where *muis* (mouse) is an existing Dutch word) learned to interpret the same ambiguous sound as /s/. In a subsequent phonetic-categorization task, the listeners exposed to ambiguous /f/-final stimuli characterized stimuli on an [ɛf-ɛs] continuum more often as an [ɛf] than the other group, thus showing a retuning of their /f/ phoneme category.

For lexically-guided retuning to occur, the availability of lexical knowledge is critical. Since non-native listeners typically have an impoverished vocabulary in comparison to native listeners, their lexical-phonological knowledge is arguably less reliable (Garcia Lecumberri et al., 2010). Secondly, non-native listeners might not have a phonetic category for the non-native sound or it might differ from the one in the native sound system (Flege, 1995). It is therefore questionable whether non-native listeners would be able to retune (non-native) phonetic categories.

Results of a previous study on lexically-guided retuning by non-native

listeners were not conclusive. Reinisch, Weber, and Mitterer (2013) demonstrated that, on the basis of Dutch L2 input, highly proficient German learners of Dutch showed retuning for ambiguous /f-s/ sounds. These sounds are however highly similar in Dutch and German. As Reinisch et al. (2013) point out, their results could be explained by L2 input guiding L1 retuning rather than retuning L2 phoneme categories since non-native phonemes that are perceived as similar to native phonemes are assimilated to these native phoneme categories (Best & Tyler, 2007; Flege, 1995). In the present study, we investigate the question whether an ambiguous sound in a non-native language can retune L2 phonetic categories, using the British English sound contrast /ɹ/-/l/. The here-presented results extend those reported in Drozdova, van Hout, and Scharenborg (2014, 2015).

While articulation of /l/ is fairly similar in Dutch and English (B. Collins & Mees, 1999), marked phonetic differences exist between British English and Dutch /r/. Realization of /r/ in Dutch depends on its position and on the speaker (e.g., Sebregts, 2015). In the onset position, uvular /ʀ/ trills or alveolar /r/ taps and trills are used while the variant closest to the British English one, the prevelar bunched approximant /ɹ/, only occurs in coda position (Mitterer, Scharenborg, & McQueen, 2013; Scobbie, Sebregts, & Stuart-Smith, 2009; Van de Velde & van Hout, 1999). British English, being non-rhotic, does not allow /ɹ/ in post-vocalic position (B. Collins & Mees, 1999). Dutch listeners thus would have to create a language-specific phonetic category for British English /ɹ/. If lexically-guided retuning is observed for Dutch listeners for the British English /ɹ/-/l/ sound contrast, this would then indicate that L2 phonetic boundaries can be retuned on the basis of ambiguous L2 input.

In the present study, Dutch non-native listeners of English were first exposed to one of two versions of a short story containing ambiguous [l/ɹ] sounds (Eisner & McQueen, 2006) and, subsequently, had to perform a phonetic-categorization task. We predict that if non-native listeners retune their L2 phonetic categories, non-native listeners in the /ɹ/-ambiguous group will show a greater proportion of /ɹ/ responses in the categorization task than listeners in the /l/-ambiguous group. To provide a baseline for the learning effect, a separate group of Dutch listeners only performed the categorization task (following, e.g., Zhang & Samuel, 2014). Moreover, since retuning has not yet been demonstrated for British English native listeners for the /l/-/ɹ/ contrast, a group of native British English listeners functioned as a control group.

2.2 Method

2.2.1 Participants

Eighty-nine native Dutch participants were recruited from the Radboud University Nijmegen subject pool of which 15 (3 males, M_{age}=23.1, SD=4.7) took part in the pretest of the stimuli, and 20 (6 males, M_{age}= 21.1, SD=2.3) in the baseline experiment. The remaining 54 participants (11 males, M_{age} =21.6, SD=2.0) participated in the main experiment. All Dutch participants possessed a 'VWO' (i.e., pre-university education) diploma, indicating a B2 or higher level of English according to the European Framework of Reference. As a control group, 47 native British English participants (9 males, M_{age}=21.0, SD=2.2) were recruited from the participant pool of the University of York, UK. All participants were paid for their participation and none reported a history of hearing or learning disorders.

2.2.2 Materials

Nineteen English words containing one /ɹ/ sound and no /l/ sounds and 19 words containing one /l/ sound and no /ɹ/ sounds with word frequencies of at least 100 per million were selected from the CELEX database (Baayen, Piepenbrock, & Gulikers, 1995). Since lexically-guided retuning is allophone-dependent (Mitterer et al., 2013), the /l/ or /ɹ/ target sound always occurred at the onset of the third or fourth syllable. The target words were embedded in a short story, which importantly, contained no other words with /l/ and /ɹ/. The story was recorded by a male native speaker of British English in three versions (normal; each /ɹ/ pronounced as /l/; each /l/ pronounced as /ɹ/). The target words were excised at the positive-going zero crossings. Two versions of each word (e.g., *memory-memoly*) were morphed with the STRAIGHT algorithm (Kawahara, Masuda-Katsuse, & De Cheveigne, 1999) to create an 11-step ambiguous word continuum where the interpretation of the ambiguous [l/ɹ] sound ranged from /l/ (step 0) to /ɹ/ (step 10). The most ambiguous variant of each individual word was chosen on the basis of a pretest with Dutch listeners. It was the step on the continuum that received approximately 50% /ɹ/ and 50% /l/ responses. This step was spliced back into the story to create two versions. In the /l/-ambiguous version, the /l/ sound in all words was replaced by the ambiguous [l/ɹ] sound

while all /ɹ/ sounds were natural. In the /ɹ/-ambiguous version, all /ɹ/ sounds were replaced with the ambiguous [l/ɹ] sound while all /l/ sounds were natural (see Appendix A for the short story).

In the phonetic- categorization task, two minimal pairs were used: *alive-arrive* and *collect-correct*. According to CELEX (Baayen et al., 1995), *alive* is the most frequent word of the *alive-arrive* pair, while *correct* is the most frequent of the *collect-correct* pair, thus reducing bias towards an /l/ or /ɹ/ interpretation in the task. The words were recorded by the same male speaker, and morphed following the procedure described above. The test phase consisted of five steps from each minimal pair: the most ambiguous item chosen on the basis of the pretest, and the two steps directly preceding and following it. For the *alive-arrive* minimal pair these were steps 3-7, and for *collect-correct* these were steps 2-6.

2.2.3 Procedure

During the exposure phase, half of the non-native and native control group participants heard the /ɹ/-ambiguous version of the story while the other half listened to the /l/-ambiguous version. In the subsequent phonetic-categorization task, participants heard the 120 test stimuli divided over four blocks. They categorized the stimuli as containing an /l/ (left button on the button box) or /ɹ/ (right button on the button box). The whole procedure took approximately twenty minutes. Participants in the baseline condition only performed the phonetic-categorization task, which lasted approximately ten minutes.

2.3 Results

In the non-native condition, 26 Dutch participants listened to the /ɹ/-ambiguous version of the story and 28 to the /l/-ambiguous version. In the native English, control condition, 23 participants listened to the /ɹ/-ambiguous version and 24 to the /l/-ambiguous version. A generalized linear mixed-effect model analysis (Baayen, Davidson, & Bates, 2008) was conducted on the responses in the phonetic-categorization task (/l/ coded as 0 or /ɹ/ coded as 1) using the logit link function. The analysis started from the model containing all predictors and all possible interactions between Exposure Condition (the critical variable), Continuum Step (the most /l/-like step was recoded as step 1, and the most /ɹ/-like

step was recoded as step 5), and Minimal Pair. Additionally, by-Subject and by-Minimal-Pair random intercepts and slopes were added to the model. Subsequently, interactions and predictors that were not significant were removed one-by-one, and each subsequent model was compared with the previous one using the likelihood ratio test. Final model was selected by comparing AIC values on the basis of likelihood ratio tests and degrees of freedom (the number of factors).

Although the analysis of the responses of the native listeners in the phonetic-categorization task did not reveal a general effect of Exposure Condition (β=- 0.67, SE=0.437, p=0.125), there was a significant interaction between Exposure Condition and Minimal Pair (β=1.296, SE=0.565, p <.05). Consequently, separate analyses were carried out for each minimal pair. Table 2.1 displays the estimates of the fixed effects and their interactions in the best-fitting model for the native listeners for the *collect-correct* (upper part) and the *alive-arrive* (lower part) minimal pair.

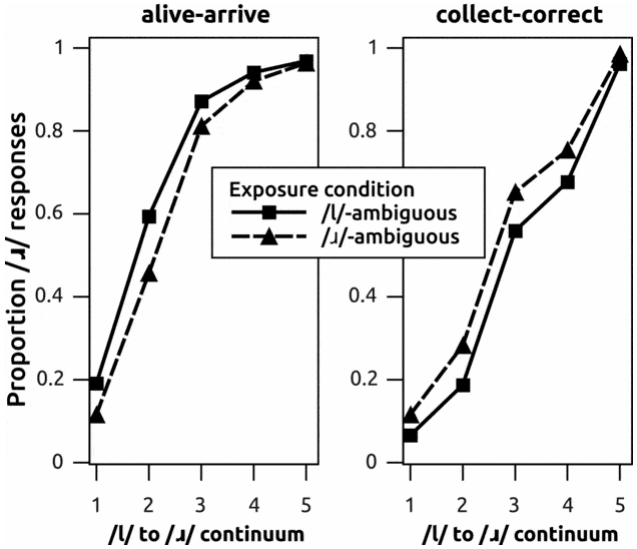

Figure 2.1: Proportion of /ɹ/ responses of the native listeners for the *alive-arrive* (left) and *collect-correct* (right) test continua. Responses of the listeners in the /ɹ/-ambiguous group are plotted with the dashed line with triangles; responses of the listeners in the /l/-ambiguous group are plotted with the solid line with squares.

Table 2.1: Fixed-effect estimates of the performance of the native listeners in the phonetic categorization task, for the minimal pairs separately

Fixed effect	β	SE	$p<$
collect-correct			
Intercept	-3.977	0.451	.001
Exposure Condition	1.363	0.602	.05
Step 2	1.663	0.339	.001
Step 3	4.486	0.368	.001
Step 4	5.253	0.378	.001
Step 5	8.265	0.491	.001
Exposure Condition x Step 2	-0.272	0.430	*ns*
Exposure Condition x Step 3	-1.016	0.460	.05
Exposure Condition x Step 4	-1.148	0.475	.05
Exposure Condition x Step 5	-0.693	0.767	*ns*
alive-arrive			
Intercept	-2.325	0.261	.001
Step 2	2.544	0.185	.001
Step 3	4.578	0.216	.001
Step 4	5.639	0.252	.001
Step 5	6.452	0.306	.001

Figure 2.1 shows the proportion of /ɹ/ responses for the 5 steps of the /l/-/ɹ/ test continuum for the /l/-ambiguous group (solid line with squares) and the /ɹ/-ambiguous group (dashed line with triangles) for the results for *alive-arrive* (left panel) and *collect-correct* (right panel). The difference between the curves of the two exposure groups, indicating the lexically-guided retuning effect, was significant only for *collect-correct* (see Table 2.1, Exposure Condition; Exposure Condition was not significant for *alive-arrive* in the final, best-fitting model, Exposure Condition in the penultimate model: β=- 0.668, SE=0.436, p=0.125). Native participants, therefore, demonstrated lexically-guided retuning for the /l/-/ɹ/ continuum, albeit only for the *collect-correct* pair.

Table 2.2 displays the estimates of the fixed effects and their interactions in the best-fitting model for the non-native listeners for both minimal pairs together. Similar to Figure 2.1, Figure 2.2 shows the pro-

portion of /ɹ/ responses of the non-native listeners for the 5 steps of the /l/-/ɹ/ continuum for the two minimal pairs separately. Crucially, the /ɹ/-ambiguous group gave significantly more /ɹ/ responses than the /l/-ambiguous group (Table 2.2, Exposure Condition), showing lexically-guided retuning.

Table 2.2: Fixed-effect estimates of performance of non-native listeners in the phonetic categorization task.

Fixed effect	β	SE	$p<$
Intercept	-2.428	0.287	.001
Exposure Condition	1.188	0.359	.001
Step 2	1.816	0.209	.001
Step 3	4.092	0.229	.001
Step 4	5.493	0.265	.001
Step 5	5.770	0.297	.001
Minimal pair	-1.003	0.324	.01
Exposure Condition x Step 2	-0.741	0.240	.01
Exposure Condition x Step 3	-1.145	0.260	.001
Exposure Condition x Step 4	-1.845	0.290	.001
Exposure Condition x Step 5	-1.606	0.352	.001
Minimal pair x Step 2	-0.101	0.238	*ns*
Minimal pair x Step 3	0.336	0.256	*ns*
Minimal pair x Step 4	0.339	0.285	*ns*
Minimal pair x Step 5	2.180	0.367	.001

To investigate whether retuning occurred for the non-native listeners for the crucial /ɹ/ category, the responses of both exposure groups were compared to the responses of the 20 non-native participants in the baseline condition (no exposure to the ambiguous sound; see dotted line with bullets in Figure 2.2). In the statistical analysis, baseline was added as a third level to the factor Exposure Condition as reference category. No significant difference was found between the baseline condition and the /l/-ambiguous group (see Table 2.3: Exposure Condition-/l/-amb). Crucially, however, the listeners in the /ɹ/-ambiguous group gave significantly more /ɹ/ responses in the phonetic-categorization task than the baseline group (Table 2.3: Exposure Condition-/ɹ/-amb). Retuning

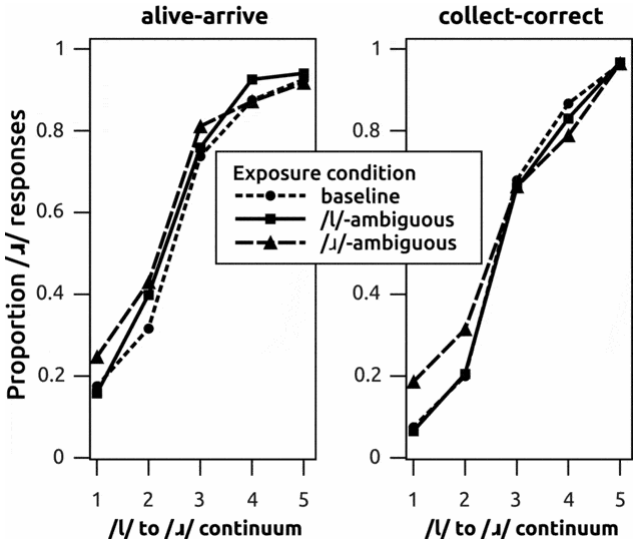

Figure 2.2: Proportion of /ɹ/ responses of the non-native listeners for the *alive-arrive* (left) and *collect-correct* (right) test continua. Responses of the listeners in the /ɹ/-ambiguous group are plotted with the dashed line with triangles; responses of the listeners in the /l/-ambiguous group are plotted with the solid line with squares; responses of the listeners in the baseline group are plotted with the dotted line with bullets

was thus observed for the group exposed to the /ɹ/-ambiguous version but not the /l/-ambiguous version of the story[1].

[1]Additional analyses investigating a possible difference in the magnitude of perceptual learning between the English and Dutch listeners showed no significant interaction between Language Group and Exposure Condition. Only for the *alive-arrive* pair the interaction between Exposure Condition and Language Group was significant (β=-1.535, SD=0.667, p<.05), which is in agreement with the earlier observed difference between the language groups regarding *alive-arrive*.

Table 2.3: Fixed-effect estimates of performance of non-native listeners in the phonetic categorization task including the baseline condition.

Fixed effect	β	*SE*	$p<$
Intercept	-2.112	0.315	.001
Exposure condition-/l/-amb	-0.285	0.390	*ns*
Exposure condition-/ɹ/-amb	0.913	0.379	.05
Step 2	1.201	0.223	.001
Step 3	3.607	0.228	.001
Step 4	4.852	0.272	.001
Step 5	5.215	0.310	.001
Minimal pair	-1.113	0.290	.001
Exposure condition-/l/-amb x Step 2	0.568	0.278	.05
Exposure condition-/ɹ/-amb x Step 2	-0.176	0.255	*ns*
Exposure condition-/l/-amb x Step 3	0.404	0.278	*ns*
Exposure condition-/ɹ/-amb x Step 3	-0.744	0.274	.05
Exposure condition-/l/-amb x Step 4	0.493	0.338	*ns*
Exposure condition-/ɹ/-amb x Step 4	-1.350	0.306	.001
Exposure condition-/l/-amb x Step 5	0.517	0.398	*ns*
Exposure condition-/ɹ/-amb x Step 5	-1.092	0.370	.01
Minimal pair x Step 2	0.032	0.208	*ns*
Minimal pair x Step 3	0.555	0.224	.05
Minimal pair x Step 4	0.671	0.249	.01
Minimal pair x Step 5	2.297	0.317	.001

2.4 Discussion and conclusions

According to the PAM-L2 model (Best & Tyler, 2007), L2 phonemes that are perceived as phonologically similar but phonetically deviant from the corresponding L1 phonemes are dissimilated from their L1 equivalent at the phonetic level, and form a separate language-specific phonetic category. This hypothesis is corroborated by the finding that French listeners of English tend to perceive British English /ɹ/ as /w/-like, despite French and English both having a phonological category for /r/ (Hallé, Best, & Levitt, 1999). French listeners in that study dissimilated English /ɹ/ from French /r/ at the phonetic level because English and French

/r/ differ phonetically. In our experiment, English /l/ is most likely assimilated with Dutch /l/, as the phonetic realization of /l/ in English and Dutch is highly similar. On the other hand, marked differences exist in the phonetic realizations of British English and Dutch /r/, similar to the French case (Hallé et al., 1999). The Dutch /r/ closest to British English /ɹ/ only occurs in coda position, where the British English /ɹ/ does not appear. Following PAM-L2, British English /ɹ/ would require a language-specific phonetic category for the Dutch listeners.

The proportion of /ɹ/ responses of the non-native listeners exposed to the /ɹ/-ambiguous version of the story was significantly larger than that of the non-native listeners in the /l/-ambiguous and the baseline groups, demonstrating retuning in non-native listeners in the /ɹ/-ambiguous condition. Three explanations seem possible for this finding: the size of the native /l/-category shrank to exclude ambiguous pronunciations, the size of the non-native /ɹ/-category widened to include ambiguous pronunciations, or a combination of both happened. Solely a retuning of the /l/-category seems to be the least plausible, as the /ɹ/-ambiguous participants were exposed to natural /l/-tokens, and were unaware that /l/-/ɹ/ was the target contrast in the study. Moreover, if exposure to an ambiguous sound in an /ɹ/-context would lead to shrinking of the /l/-category, one would also expect a reduction in the size of the /ɹ/-category for the listeners in the /l/-ambiguous group. This was not found. So, even if exposure to the ambiguous sound caused a reduction in the size of the /l/-category, the size of the /ɹ/-category needs to increase to account for our data.

In contrast to what is typically found in native listeners (e.g., compare Figures 2.1 and 2.2), the non-native retuning seems to be concentrated mostly on the /l/-side of the continuum. Potentially, native listeners have a better developed /ɹ/-category than non-native listeners, helping them to flexibly adjust category boundaries when faced with ambiguous pronunciations. The arguably less well-developed /ɹ/-category of the non-natives might result in a retuning that is more bound to specific rather than a range of ambiguous steps. Taken together, we conclude that the observed retuning effect seems to suggest that L2 listeners can, in addition to their L1, also retune their L2 phonetic categories, albeit perhaps in a 'narrower' sense than native listeners.

The responses in the phonetic-categorization task of the non-native /l/-ambiguous group did not differ significantly from those of the non-native baseline group. This asymmetry in lexically-guided retuning has

previously been demonstrated for the /s/-/f/ contrast (Eisner & McQueen, 2006; Norris et al., 2003; Zhang & Samuel, 2014), where the /s/-ambiguous group experienced a stronger retuning than the /f/- ambiguous group. Zhang and Samuel (2014) argued that the frication cue of /f/ is weaker than that of /s/ and therefore more susceptible to variation, which in turn would block retuning. We want to add a complimentary interpretation: the acoustic variation for /ɹ/ both in British English and Dutch is high both in allophonic variation and between speakers, higher than that of /l/. Like /ɹ/, /s/ has large inter-speaker variation (Dart, 1991, 1998), more so than /f/ (Gordon, Barthmaier, & Sands, 2002), suggesting that the acoustics of /s/ are inherently more variable than those of /f/. Taken together, potentially, listeners more easily retune phoneme categories of sounds which are acoustically variable, after exposure to artificially-induced variation, as they are used to hearing this variation and adapting these phoneme categories. This explanation is in agreement with studies on talker variability (Bradlow & Bent, 2008; Clopper & Pisoni, 2004) which show that highly variable training stimuli (e.g., exposure to different voices) promote perceptual learning. Whether indeed sounds with high(er) inherent acoustic variability are more prone to retuning than more stable sounds is an interesting question for future research.

Surprisingly, lexically-guided retuning for the native listeners only occurred for the *collect-correct* pair. Potentially, the steps used for the *alive-arrive* pair were not well positioned for native listeners as the ambiguous steps were chosen on the basis of a pre-test with non-native listeners. Post-hoc acoustic analyses indeed revealed that the first step of the *alive-arrive* continuum was more /ɹ/-like than the first step of the *collect-correct* continuum (see Drozdova, van Hout, & Scharenborg, 2014, for more discussion). This slight /ɹ/-bias could possibly reduce the perceptual learning effect.

To summarize, our results suggest that non-native listeners are able to retune their non-native phonetic boundaries. This suggests that the mechanisms underlying lexically-guided perceptual learning in non-native listening correspond to those observed in native listening (Norris et al., 2003; Samuel & Kraljic, 2009), although in somewhat narrower sense, and that non-native listeners enjoy a similar remarkable flexibility at the prelexical/phonetic level which has previously been associated with native listening (Cutler, 2012).

CHAPTER 3

The effect of intermittent noise on lexically-guided perceptual learning in native and non-native listening

This Chapter is based on
Drozdova, P., van Hout, R., & Scharenborg, O. (2017). The effect of intermittent noise on lexically-guided perceptual learning in native and non-native listening. Revision of the manuscript submitted for publication.

Abstract

There is ample evidence that both native and non-native listeners deal with speech variation by quickly tuning into a speaker, even when pronunciations are ambiguous. Noise in the speech signal has previously been shown to change the competition process in word recognition, it increases the number of candidate words competing for recognition and slows down the recognition process. Given the role of lexical information in inducing the adjustment of phonetic categories, the present study investigates whether intermittent noise interferes with lexically-guided perceptual learning in native and non-native listening. Native English and Dutch listeners were exposed to a short story in English, where either all /l/ or all /ɹ/ sounds were replaced by an ambiguous sound half-way between /l/ and /ɹ/. Although both the native and the non-native groups exposed to the short story in the clean condition demonstrated lexically-guided perceptual learning, non-native listeners showed only incipient or no lexically-guided perceptual learning when intermittent noise was added to the short story, even though the noise never occurred on the critical items. We argue that this native-non-native difference can be explained by the interplay of the effects of noise and non-nativeness on the lexical competition process during spoken-word recognition.

3.1 Introduction

There is enormous variation among speakers in how they produce sounds and words. This is due to differences in the speakers' accent, dialect, speaking style, and idiosyncrasies of their vocal tract or, for instance, because the speaker has a speech impediment. There is ample evidence that listeners deal with this variation by quickly tuning into a speaker, even when pronunciations are ambiguous (Norris et al., 2003). In order to do so, listeners can use their lexical knowledge (see Samuel & Kraljic, 2009 for an overview). The mechanism through which adaptation occurs is termed lexically-guided perceptual learning or lexical retuning (Norris et al., 2003). It has been argued to aid listeners in adapting to unfamiliar speakers producing certain sounds in an unusual way (e.g., Norris et al., 2003; Reinisch & Holt, 2014), and in facilitating the recognition of future speech input containing the same ambiguous sound patterns (McQueen, Cutler, & Norris, 2006).

Lexically-guided perceptual learning was first demonstrated by Norris and colleagues (2003). In their study, Dutch listeners were exposed to words in Dutch with an ambiguous sound half-way between /f/ and /s/, denoted as [f/s], in a lexical decision task. One group of listeners heard /f/-final words where the final /f/ sound was replaced by the ambiguous sound (e.g., *witlo[f/s]* - chicory). These listeners learned to interpret this ambiguous sound as an /f/, since the word *witlof* is an existing Dutch word while *witlos* is not. The other group of listeners was exposed to /s/-final words where the final /s/ was replaced by the ambiguous [f/s] sound. These listeners learned to interpret the ambiguous [f/s] sound as an /s/, as the /s/-interpretation of the stimulus is an existing Dutch word while the /f/-interpretation is not (e.g., *baa[f/s]*, where *baas* is a Dutch word (boss) and *baaf* is not). Retuning revealed itself in a subsequent phonetic categorization task, where the listeners exposed to the ambiguous items in /f/-final words interpreted stimuli on an [ɛf-ɛs] continuum more often as an /ɛf/ than the listeners exposed to the ambiguous /s/-final words. Exposure to an ambiguous sound triggers a temporary change in listeners' phonetic representations (Clarke-Davidson et al., 2008). Lexically-guided perceptual learning generalizes to words that have not been presented earlier (McQueen, Cutler, & Norris, 2006), so that, e.g., Dutch adults interpret the previously unheard word *lɔ[f/s]* as *lof* (praise) or *los* (loose) depending on their previous exposure to *witlo[f/s]* or *baa[f/s]*, respectively (Scharenborg, Weber, &

Janse, 2015). Generalization of learning to words not present in the exposure phase strongly suggests that adjustment occurs at the prelexical level of processing (McQueen, Cutler, & Norris, 2006).

Norris et al. (2003) showed that lexically-guided perceptual learning only occurs when the ambiguous sound is included in an existing word but not when the ambiguous sound is embedded in a non-word, and concluded that listeners adjust their phonetic category boundaries only when their lexical knowledge can be exploited to interpret ambiguous stimuli. Cutler et al. (2008) extended this proposition showing that ambiguous sounds in non-words can also induce phonetic category retuning, but only when they are part of a legal sequence of phonemes in the listener's native language. Jesse and McQueen (2011) further studied the role of lexical information for lexically-guided perceptual learning. They demonstrated that no learning occurs in native listening when ambiguous sounds are located at the start of words, and argued that in order for lexically-guided perceptual learning to occur, lexical knowledge should be available quickly and should be reliable enough to guide retuning. Although the words containing the ambiguous sounds were recognized as words (80% acceptance rate on the ambiguous items in the lexical decision task which was used in the exposure phase), the disambiguating information was available too late relative to the position of the ambiguous sound at the start of the word for lexically-guided perceptual learning to occur. This again shows the importance of lexical information for lexically-guided perceptual learning, and suggests that lexical competition should be resolved early enough to trigger retuning.

Because of the essential role of lexical information in lexically-guided perceptual learning, non-native listeners, who have less stable, detailed, and abstract lexical knowledge than native listeners (Garcia Lecumberri et al., 2010), might possibly be hampered in adapting to ambiguous sounds in non-native speech. Moreover, phonetic categories and contrasts present in the non-native language might be absent or realized differently from those in the native language of the listener (Flege, 1995), which could result in failure to recognize the ambiguous sound or not treating it as ambiguous enough to induce retuning. However, highly proficient non-native listeners do show lexically-guided perceptual learning, and are able to retune both native (Reinisch et al., 2013) and non-native (Drozdova, van Hout, & Scharenborg, 2016) phonetic categories. Both lexical knowledge and the phoneme categories of the non-native listeners thus seem to be defined well enough to allow phonetic category retuning

to occur. Both native and non-native phonetic category representations are thus rather flexible, at least when non-native listeners are relatively proficient.

There are, however, clear bounds to this flexibility. Samuel and Kraljic (2009) argue that retuning is blocked when variation in the signal can be attributed to speaker-external factors. Kraljic, Brennan, and Samuel (2008) demonstrated that acoustic deviations due to context-dependent variability, e.g., caused by a certain dialect (e.g., the pronunciation of /s/ as /ʃ/ when followed by /tr/ in Philadelphia English), prohibited adaptation in native listening. Similarly, no retuning emerges when the ambiguity in the signal is caused by a pen in the mouth of the speaker (Kraljic, Samuel, & Brennan, 2008). Another speaker-external factor blocking lexically-guided perceptual learning was found to be the presence of background noise. Zhang and Samuel (2014) added signal-correlated noise to their stimuli in the exposure phase, masking both the carrier sentences and the critical lexical items, but not the ambiguous sound (a sound between /f/ and /s/). In contrast to listeners who performed the same task in clean, no lexically-guided perceptual learning was observed for listeners exposed to the stimuli masked by noise. Zhang and Samuel (2014) hypothesized that when the speech signal is noisy and hence more variable, native listeners do not treat the ambiguous sound as a reliable cue to trigger retuning.

The presence of noise in the speech signal has also been found to change the dynamics of phonological competition in native listeners (Ben-David et al., 2011; Brouwer & Bradlow, 2011, 2016; Hintz & Scharenborg, 2016; McQueen & Huettig, 2012). McQueen and Huettig (2012) found that intermittent noise elsewhere in the signal made the native listeners in their eye-tracking study, subconsciously, less confident about which words they heard, and hypothesized that the presence of intermittent noise increased listeners' expectation of a distortion occurring. This increased expectation of an ambiguity occurring was shown to lead to a change in the lexical competition process which showed itself as an increase in the number of looks at the rhyme competitors and decrease in the number of looks at the onset competitors in comparison to the clean listening conditions. Moreover, the presence of noise has been shown to increase the time listeners need to resolve competition in spoken word recognition (e.g., Ben-David et al., 2011; Brouwer & Bradlow, 2011, 2016). This slowing down is due to an increase in the number of candidate words competing for recognition when noise is present (Scharenborg

et al., 2017), a longer activation of the candidate words in the memory of the listeners (Brouwer & Bradlow, 2011), and a reduced activation of the candidate words (Hintz & Scharenborg, 2016). Relatedly, an eye-tracking study with cochlear implant (CI) users (Farris-Trimble, McMurray, Cigrand, & Tomblin, 2014) found differences in the degree of peak and late competitor activations between CI users and a CI simulation group of normal hearing participants. They hypothesized that, similar to the participants of McQueen and Huettig (2012), CI users keep competitors active in memory longer as they are accustomed to degraded input, and that consequently this delays commitment to lexical items. Listeners are thus able to flexibly adjust their interpretation of acoustic information and consequently their spoken-word recognition processes as listening conditions change (see also Brouwer, Mitterer, & Huettig, 2012).

Listening in noise is typically found to be more challenging for non-native than for native listeners (e.g., Mayo et al., 1997; Rogers, Lister, Febo, Besing, & Abrams, 2006; see for a review Garcia Lecumberri et al., 2010). Non-native listeners, therefore, may provide an eminent testing ground for establishing the interaction of two potentially crucial factors in lexically-guided perceptual learning: characteristics of the signal and lexical knowledge available to the listeners. When the speech signal contains background noise, the phonological match between the target word and the activated words decreases (see Garcia Lecumberri et al., 2010), this results in an increase of the number of candidate words compared to clean listening conditions that is even larger than that for native listeners (Scharenborg et al., 2017).

The present study investigates the effect of intermittent noise on lexically-guided perceptual learning in native and non-native listening. Given the effect of (intermittent) noise on interpreting lexical information in the speech signal, lexically-guided perceptual learning might be impeded in noise even when this noise is intermittent and never occurs on the critical words, and more so for non-native than for native listeners. Two experiments were conducted to investigate this hypothesis: in the first experiment, native listeners of English were auditorily presented with one of two versions of a short story in English in which for all words with an /l/ or /ɹ/ sound, the /l/ or /ɹ/ sound was replaced by an ambiguous [l/ɹ]. The stories were presented either in a clean version or in a version with intermittent noise. In the noise version of the short story, parts of the speech stimuli were masked with noise, while,

crucially, words containing the target ambiguous sound were left intact. In the second experiment, Dutch non-native listeners of English were exposed to one of the two versions of the same short story as the native listeners, again either in clean or in noise listening conditions. Articulation of /l/ is similar in Dutch and English (B. Collins & Mees, 1999), while British English prevelar bunched approximant /ɹ/ only occurs in Dutch in coda position (Mitterer et al., 2013; Scobbie et al., 2009; Van de Velde & van Hout, 1999), where it never occurs in English. Dutch listeners would thus have to create a language-specific phonetic category for British English /ɹ/ (Drozdova et al., 2016). After listening to the short story, all participants performed a phonetic categorization task.

3.2 Method

Following the standard procedure for lexically-guided perceptual learning studies (e.g., Norris et al., 2003; Scharenborg et al., 2015; Zhang & Samuel, 2014), the main experiment consisted of two parts: an exposure phase and a test phase. The exposure phase consisted of a short story (similar to Drozdova et al., 2016; Eisner & McQueen, 2006) with a between-subject manipulation (see Appendix A and B). Half of the participants listened to the story where all /l/ sounds were replaced by an ambiguous [l/ɹ] sound, while the other half of the participants listened to the same story where all /ɹ/ sounds were replaced by the ambiguous [l/ɹ] sound. Participants were randomly assigned to one of the two versions of the short story. During the test phase, all participants had to perform a phonetic categorization task, followed by five comprehension questions about the short story for the non-native group of participants (see Appendix C). To obtain a measure of the lexical proficiency in English of the non-native listeners, LexTALE (Lexical Test for Advanced Learners of English: Lemhöfer & Broersma, 2012) was administered to the non-native listeners. LexTALE is an unspeeded lexical decision task in which participants are exposed to 60 items and have to decide upon the presentation of each item, whether it is an existing word in English or not.

3.2.1 Participants

One hundred nine native English listeners (24 males, M_{age} = 20.7, SD = 1.9), recruited from the Psychology Electronic Experiment Booking

system of the Department of Psychology of the University of York, participated in the native versions of the experiment. Note that the data of the native participants in the clean listening conditions were acquired and have also been analyzed in the context of a different project (see Drozdova et al., 2016 and Chapter 2).

Eighty native Dutch participants (8 males, M_{age} =21.6, SD = 2.2) were recruited from the Radboud University Nijmegen subject pool, and participated in the non-native versions of the experiment. An overview of the number of participants for each listening and exposure condition per language is presented in Table 3.1. Prior to the experiment, all native and non-native participants had to fill in a questionnaire with questions regarding any hearing or learning disorders and possible difficulties in hearing in the presence of background noise. Only participants without learning or hearing disorders were included in the experiment. All participants participated in only one version of the experiment.

Table 3.1: The number of participants in each experiment assigned to the /l/-ambiguous (/l/-amb.) or the /ɹ/-ambiguous (/ɹ/-amb.) version of the story in clean and noisy conditions.

	Clean		Noise	
Listeners	/ɹ/-amb.	/l/-amb.	/ɹ/-amb.	/l/-amb.
Native	28	26	29	26
Non-native	20	20	20	20

Additionally, 15 native Dutch participants (3 males, M_{age} = 23.1, SD= 4.7) took part in a pretest of the stimuli, and another, separate group of eight native Dutch participants (M_{age}= 22, SD = 2.8) took part in a pilot study to determine the appropriate length of the noise fragments in the noise condition. None of these participants participated in the main experiments. All participants received a monetary reward for their participation, and signed a consent form prior to the experiment.

3.2.2 Exposure phase: clean

The story used in the exposure phase was created in the context of a previous experiment (Drozdova et al., 2016). It included 19 words containing one /l/ sound and no /ɹ/ sounds, and 19 words containing one

/ɹ/ sound and no /l/ sounds. The words were chosen from the CELEX database (Baayen et al., 1995) and had frequencies of at least 100 per million. Since lexically-guided perceptual learning is impeded when listeners hear standard pronunciations of the target sound from the same speaker (Kraljic & Samuel, 2011) no words in the story other than the target words contained /l/ or /ɹ/. As retuning does not transfer to other allophones of the same sound (Mitterer et al., 2013), we insured that /l/ or /ɹ/ occurred in the same position for all target words, i.e., at the onset of the third or fourth syllable (except for one word: *Internet*). The final version of the story consisted of 333 words, of which 38 were critical items (see Appendix A for the short story). The total duration of the short story was 2.21 minutes.

The story was recorded by a male native speaker of British English from South West England in a sound-attenuated booth with a Sennheiser ME 64 microphone at the sampling frequency of 44100 Hz. In order to obtain the ambiguous sound between /l/ and /ɹ/, the story was recorded in three versions: in one version all words were pronounced in the natural way, in the second version all words containing an /l/ sound were pronounced with an /ɹ/ sound (e.g., *accumurated*), in the third version all /ɹ/ sounds were substituted with /l/ sounds (e.g., *wondeling*). The words were then excised at the positive-going zero crossings from each version of the short story and zero-padded with 25 ms silence at the onset and the offset using Praat (Boersma & Weenink, 2009). The pitch contours of the two items from each pair (e.g., *memory-memoly*) were equalized and, following the procedure described by Scharenborg and Janse (2013), were morphed with the STRAIGHT algorithm (Kawahara et al., 1999). STRAIGHT first decomposes the input files into source parameters and spectral parameters, and subsequently removes pitch information, while keeping frequency information. In order to keep coarticulatory information of upcoming /l/ and /ɹ/ in the syllable preceding the critical sound available to the listener, whole words were morphed rather than separate sounds. As a result of morphing the item-pairs, an 11-step continuum was created where step 0 was the most /l/-like sound and step 11 the most /ɹ/-like.

To determine the most ambiguous step between /l/ and /ɹ/, a pretest with 15 Dutch listeners was conducted. The pre-test consisted of a phonetic categorization task, where listeners had to decide whether they heard an item with an /l/ or an /ɹ/ sound by pressing the corresponding button on the button box. Participants listened to different steps of

the continuum, i.e., steps 1, 3, 5, 7, 9. The left button of the button box corresponded to the item containing /l/, whereas the right button corresponded to the item with an /ɹ/ sound. The two possible answers were also presented on the computer screen with the /l/-reading of the stimulus on the left side of the screen and the /ɹ/-reading on the right side of the screen. So, in half of the trials, the /l/ answer was a word and the /ɹ/ answer a non-word and in half of the trials the /ɹ/ answer was a word and the /l/ answer was a non-word. Participants categorized five steps of each target (38 words) or test word (4 words: see subsection Test Phase). Each step of the continuum was presented twice to the participants. Participants categorized 400 items in total.

The proportions of /l/ and /ɹ/ responses for the test items were calculated. The step on the continuum that received approximately 50% of both responses was chosen as the most ambiguous one. If the 50% point occurred in between two test steps, the step in between was chosen. The most ambiguous step was determined individually for each word and then spliced back to the corresponding version of the short story. Two versions of the story were created: in one version all words containing an /l/ sound were replaced by the ambiguous [l/ɹ] sound, while all /ɹ/ sounds remained natural; in the second version all words containing an /ɹ/ sound were replaced by the ambiguous [l/ɹ] sound while all /l/ sounds remained natural.

3.2.3 Exposure phase: noise

For the experiments in the noise condition, speech-shaped noise was added to the short story. Noise at a signal-to-noise ratio (SNR) of 0 dB was automatically added to fragments of the story using a Praat (Boersma & Weenink, 2009) script. First, boundaries were manually placed in the signal on the positive zero crossings in Praat. The fragments that were to be masked were marked with an X on the tier. These fragments were one to four words long. A Praat script then placed a random chunk of the noise signal on the marked part of the audio file. Before adding noise the audio file was down-sampled to 16000 Hz to match the sampling frequency of the noise file.

For lexically-guided perceptual learning to occur, listeners need to be able to comprehend the story, hence a SNR was needed that challenged listening but did not severely impair recognition accuracy. The SNR was chosen on the basis of a study by Scharenborg et al. (2017). In this study,

Dutch non-native listeners of English had an average recognition accuracy of 83.8% for English words partially embedded in speech-shaped noise at an SNR of 0 dB. This was deemed an SNR that fit our criteria. Following McQueen and Huettig (2012) noise was placed on several fragments of the story, so that at least one word, but typically two words, preceding and typically at least one word following the critical word was in the clean. The length of the noise fragments was determined on the basis of a pre-test with eight native Dutch listeners. After listening to the partially-masked story, participants had to answer five short questions to check their comprehension of the story. All eight participants answered two to four comprehension questions correctly (M = 3.25), which confirmed that the presence of noise made listening challenging but did not severely harm comprehension. None of the participants in the pre-test participated in the main experiment. For the noisified version of the short story see Appendix B.

3.2.4 Test phase

The test phase consisted of a phonetic categorization task. Two minimal pairs, not present in the target story, were used: *collect-correct* and *alive-arrive*. To avoid a bias towards either the /l/ or the /ɹ/ interpretation of the ambiguous stimuli, the two minimal pairs had an opposite pattern of word frequency, with an /l/ word being more frequent for the *alive-arrive* pair (1135 per million for *alive* and 157 per million for *arrive*) and the /ɹ/ word being more frequent for *collect-correct* (117 per million for *collect* and 804 per million for *correct*). The words were recorded by the same speaker who recorded the short story. The two members of each word pair were subsequently morphed together using the procedure described in the previous subsections. The two created continua were included in the pre-test together with the items from the short story. The test phase in the experiment included five steps from each of the two continua: the most ambiguous between /l/ and /ɹ/ step determined on the basis of the pre-test, and two steps before and after it. For the *alive-arrive* minimal pair these were steps 3-7, and for *collect-correct* these were steps 2-6.

3.2.5 Procedure

All participants were tested individually in a sound-proof booth. Prior to the experiment they filled in a consent form and a short question-

naire containing questions about their age, education, and language background. Subsequently, participants were given verbal instructions about the upcoming tasks. Additionally, they saw an instruction on the computer screen informing them that they would be listening to a short story in English. The short story was presented to the listeners binaurally through headphones. Once participants finished listening to the story, a message appeared on the screen indicating that they had to press a button on the button box to proceed to the next task. When participants pressed the button, instructions for the test phase of the experiment appeared on the screen.

The test phase was in the form of a phonetic categorization task where participants had to press a button on the button box to indicate which item (*alive* or *arrive*; *collect* or *correct*) they had just heard. The left button on the button box corresponded to the item with the /l/ sound, while the right button on the button box corresponded to the item with the /ɹ/ sound.

Since the provided testing booth for the native experiments at the University of York was not equipped with a button box, the experiment was reprogrammed such that "z" on the keyboard corresponded to the left button on the button box and "m" corresponded to the right button. To aid the participants, items were also visually presented on the computer screen. Test stimuli were divided over four blocks, with a self-paced pause after each block. Each block consisted of the five steps of each minimal pair presented three times in a random order. Participants thus listened to 120 test items. After completing the phonetic categorization task, participants in the non-native group had to answer the five comprehension questions. Exposure and test phases were followed by LexTALE.

3.3 Results

To investigate the effect of the presence of background noise on the lexically-guided perceptual learning in native and non-native listening, the responses of the listeners in the phonetic categorization task were analyzed using mixed effects logistic regression. In order to investigate lexically-guided perceptual learning, it is important that participants included in the analyses are actually able to discriminate the tokens on the continuum (see e.g., Norris et al., 2003; Zhang & Samuel, 2014). We

assumed that participants not able to discriminate /ɹ/ and /l/ sounds would hear /ɹ/ in more than 50% of the cases already on the first most /l/-like step of the continuum, or /l/ in more than 50% of the cases on the last, most /ɹ/-like step of the continuum. Moreover, participants should demonstrate a significant increase between the first and the last step of the continuum (see Zhang & Samuel, 2014 on an additional discussion on the selection of participants in lexically-guided perceptual learning studies). The three criteria that were applied were (tested using a binomial distribution, $\alpha = .05$, one-sided):

1. seven or more /l/ responses on the first step of the continuum;
2. seven or more /ɹ/ responses on the last step of the continuum;
3. difference of at least 5 between the first and the last step of the continuum

On the basis of these criteria, 11 participants were excluded from the main analysis: one non-native listener from the clean /l/-exposure group, two native listeners from the clean /l/-exposure group, five native listeners from the clean /ɹ/-exposure group and three native listeners from the noise /ɹ/-exposure group. In this way, 79 non-native and 99 native listeners were included in the analyses. Note that a previous study (Drozdova et al., 2016) showed that the learning effects were dependent on the word pair: no lexically-guided perceptual learning effect was observed in native listening for the *alive-arrive* pair. Consequently, the analyses presented here only include the *correct-collect* word pair. We will come back to this point in the Discussion section.

All analyses were performed in R (version 3.0.2), using glmer (package lme4) with the optimizer set to BOBYQA (Powell, 2009). The dependent variable was the number of /ɹ/ responses. To that end, all /l/ responses of the participants were coded as 0 and all /ɹ/ responses as 1. We started from the analysis including both the native and the non-native listener groups in both listening conditions (clean and noise) in an overall model containing all predictors: Exposure Condition (/ɹ/-ambiguous or /l/-ambiguous version of the short story), Noise (whether the story was presented in clean or in noise), Step on the continuum (the most /l/-like step was recoded as step 1, and the most /ɹ/-like step was recoded as step 5; Step 1 was chosen as a reference point), Language (whether the participant was a native or a non-native listener), and all possible four-,

three- and two-way interactions between them. Step on the continuum was included as a categorical variable and Subject was included as a random factor. A backward selection procedure was applied, in which interactions and predictors that were not significant at the 5% level were one-by-one removed from the model, starting with the least significant interactions. Each change in the fixed effect structure was evaluated by inspecting the likelihood ratio changes with the anova function. Table 3.2 gives the estimates of the fixed effects and their interactions of the best-fitting model.

Table 3.2: Fixed-effect estimates of the cross-linguistic analysis in the phonetic categorization task

Fixed effect	β	SE	p
Intercept	-3.565	0.310	<.001
Exposure Condition	1.208	0.291	<.001
Step 2	1.417	0.176	<.001
Step 3	4.788	0.181	<.001
Step 4	5.004	0.187	<.001
Step 5	7.688	0.243	<.001
Noise	-0.543	0.341	0.111
Language	-0.236	0.329	0.473
Exposure Condition x Step 2	-0.037	0.220	0.866
Exposure Condition x Step 3	-0.643	0.228	0.005
Exposure Condition x Step 4	-0.820	0.236	0.001
Exposure Condition x Step 5	-0.808	0.337	0.016
Noise x Language	0.972	0.458	0.034

As Table 3.2 shows, Exposure Condition was a significant predictor of the number of /ɪ/ responses, with listeners exposed to the short story with /ɪ/-ambiguous sounds giving more /ɪ/ responses in the phonetic categorization task than the other group. The listeners thus showed a perceptual learning effect. This effect is dependent on the Step involved (see the interaction of Step and Exposure Condition in Table 3.2). Moreover, the analysis revealed a significant interaction between Language and Noise, suggesting that there were differences between the responses of native and non-native listeners in the phonetic categorization task

modulated by the presence of noise. However, no significant 3-way interaction between Noise, Language and Exposure Condition was observed, which would have indicated differences in the role of noise in lexically-guided perceptual learning for native and non-native listeners. At the same time, there were differences in the testing conditions (equipment, location) between the native and non-native listeners, and there might be differences in error patterns or standard deviations between the two groups due to the increased effect of noise on speech processing in non-native listeners compared to native listeners (e.g., Garcia Lecumberri et al., 2010; Scharenborg et al., 2017). Consequently, additional, separate analyses were carried out for the native and the non-native listener groups to make sure that there were indeed no differences in the emergence of the lexically-guided perceptual learning effect between the two language groups.

Figures 3.1 (native listeners) and 3.2 (non-native listeners) show the log odds for choosing an /ɹ/ response for the *collect-correct* pair in the phonetic categorization task for each exposure condition for the clean (in the left panels) and the noise conditions (the right panels) separately. The log odds are calculated on the basis of the generalized linear mixed effects model, including Step on the continuum, Exposure Condition, and the interaction between them, and Subject as a random factor. Responses of the participants who were exposed to the /ɹ/-ambiguous version of the story are represented with the black dashed lines with bullets. Responses of the participants exposed to the /l/-ambiguous version of the story are shown with the gray solid lines with squares.

3.3.1 Native listeners

Responses of the native listeners were analyzed in the same way as described in the previous section (but excluding the Language factor). Similar to the main analysis, we expected to find the difference between the two exposure groups, which would manifest itself as a significant effect of Exposure Condition. Note, that the responses of the 47 listeners in the clean listening conditions were analyzed previously in Chapter 2 Drozdova et al. (2016) and reanalyzed for the present study. The estimates for the best-fitting model for the native listeners are presented in Table 3.3.

As shown in Table 3.3, irrespective of whether the native listeners were exposed to the short story in clean or in noise, no significant main

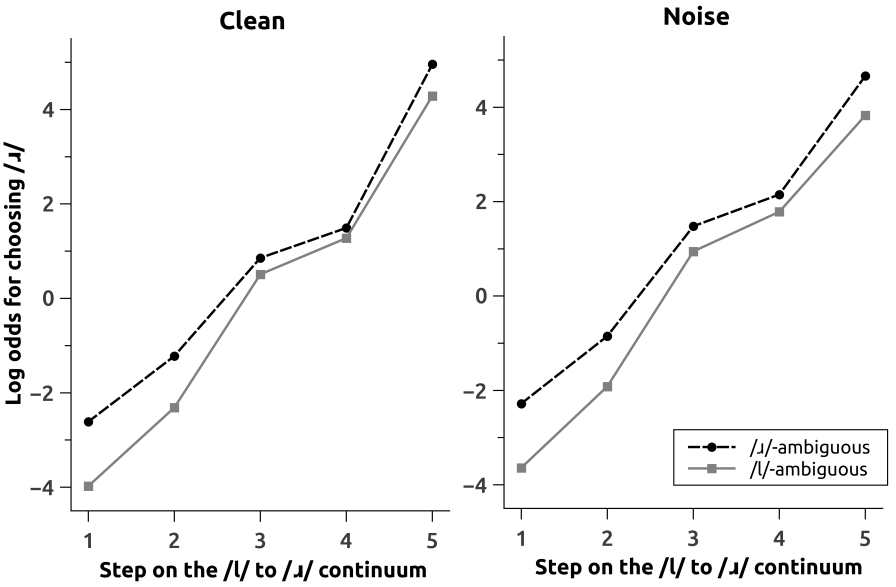

Figure 3.1: Log odds for choosing /ɹ/-responses for the native listeners for the *collect-correct* test continuum in clean (left panel) and intermittent noise (right panel) listening conditions.

effect of Noise or Noise in interaction with Step or Exposure Condition came out. In both listening conditions, those listeners exposed to the /ɹ/-ambiguous version of the story gave significantly more /ɹ/ responses in the phonetic categorization task than those listeners exposed to the /l/-ambiguous version (Exposure Condition factor). Moreover, significant interactions between Exposure Condition and Step of the continuum were observed for Steps 3 and 4, indicating that the magnitude of learning varied depending on the step of the continuum: the difference between /l/-ambiguous and /ɹ/-ambiguous exposure groups was the largest on the first, second and the last steps of the continuum and was significantly smaller on the third and the fourth continuum step than on the first step.

Table 3.3: Fixed-effect estimates of the performance of the native listeners in the phonetic categorization task

Fixed effect	β	SE	p
Intercept	-3.794	0.454	<.001
Exposure Condition	1.345	0.399	0.001
Step 2	1.695	0.237	<.001
Step 3	4.539	0.250	<.001
Step 4	5.336	0.259	<.001
Step 5	7.894	0.334	<.001
Exposure Condition x Step 2	-0.278	0.295	0.346
Exposure Condition x Step 3	-0.917	0.312	0.003
Exposure Condition x Step 4	-1.064	0.325	0.001
Exposure Condition x Step 5	-0.658	0.489	0.175

3.3.2 Non-native listeners

The estimates of the parameters included in the final model for the non-native listeners for both listening conditions together are presented in Table 3.4. Results of the non-native listeners exposed to the short story in the clean are shown in the left panel of Figure 3.2, while the results of the listeners exposed to the short story in noise are shown in the right panel of Figure 3.2.

Similar to the native listeners, non-native listeners demonstrated lexically-guided perceptual learning (significant effect of Exposure Condition, moderated by the continuum step). Although the interaction between the last step of the continuum and Exposure Condition was only marginally significant, its removal from the model significantly decreased the model fit. Importantly however, different from the native listeners' results, noise was a significant predictor of the number of /ɹ/ responses in the phonetic categorization task for the non-natives and the effect of noise was moderated by the step of the continuum. The presence of Noise in the final model for the responses of the non-native listeners and its absence from the model for the native listeners explains the significant interaction between Noise and Language in the general analysis. Moreover, it demonstrates that non-native, but not native listeners, differed in their response patterns depending on the listening conditions. To bet-

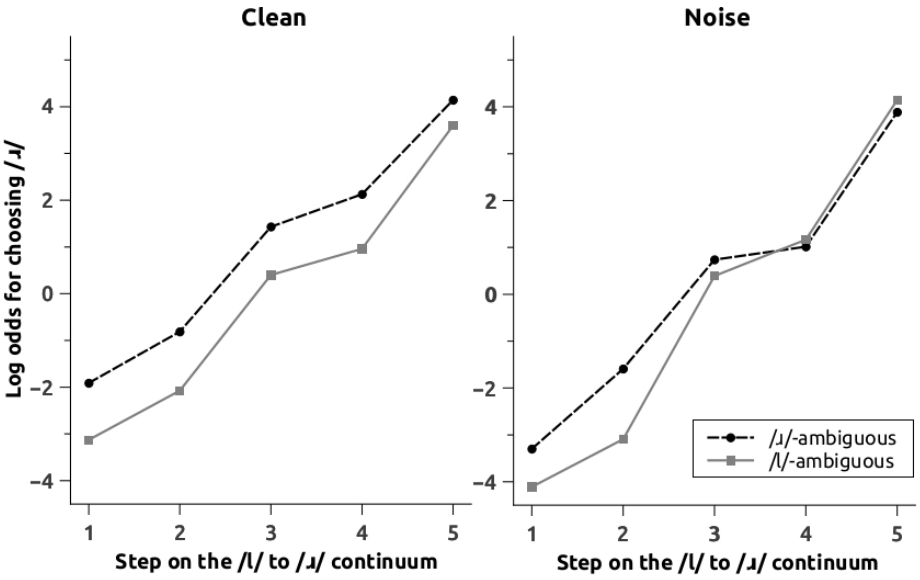

Figure 3.2: Log odds for choosing /ɹ/-responses for the non-native listeners for the *collect-correct* test continuum in clean (left panel) and intermittent noise (right panel) listening conditions.

ter understand the way the explanatory factors have an impact on each other depending on the Noise condition, the non-native listeners were analyzed separately for clean and noise in the final set of analyses.

3.3.3 Non-native listeners: clean

The estimates of the parameters that were included in the model for the non-native listeners in the clean listening condition are presented in Table 3.5.

As shown in Table 3.5 by the significant effect of Exposure Condition, non-native listeners demonstrated lexically-guided perceptual learning when they were presented with the ambiguous sounds in the clean version of the short story. The non-native listeners exposed to the /ɹ/-ambiguous version of the story gave significantly more /ɹ/ responses in the phonetic categorization task than the listeners exposed to the /l/-ambiguous version. The difference between the two exposure conditions

Table 3.4: Fixed-effect estimates of the performance of the non-native listeners in the phonetic categorization task

Fixed effect	β	SE	p
Intercept	-3.001	0.383	<.001
Exposure Condition	0.975	0.431	0.023
Step 2	0.861	0.304	0.005
Step 3	3.568	0.298	<.001
Step 4	4.267	0.307	<.001
Step 5	6.787	0.411	<.001
Noise	-1.150	0.431	0.008
Exposure Condition x Step 2	0.342	0.335	0.308
Exposure Condition x Step 3	-0.288	0.337	0.394
Exposure Condition x Step 4	-0.482	0.347	0.164
Exposure Condition x Step 5	-0.824	0.474	0.082
Noise x Step 2	0.374	0.332	0.260
Noise x Step 3	0.797	0.338	0.018
Noise x Step 4	0.702	0.347	0.043
Noise x Step 5	1.279	0.475	0.007

Table 3.5: Fixed-effect estimates of the performance of the non-native listeners in the clean listening condition in the phonetic categorization task

Fixed effect	β	SE	p
Intercept	-3.070	0.390	<.001
Exposure Condition	1.118	0.488	0.022
Step 2	1.079	0.209	0.003
Step 3	3.436	0.220	<.001
Step 4	4.054	0.229	<.001
Step 5	6.429	0.320	<.001

was present on all steps of the /l/ to /ɹ/ continuum, and there was no interaction between Steps on the continuum and Exposure Condition.

3.3.4 Non-native listeners: noise

The right panel of Figure 3.2 shows that for the first three steps, listeners from the /ɹ/-exposure group gave more /ɹ/ responses than the /l/ exposure group, while this pattern changed on the fourth step of the continuum. The best model contains the Exposure condition and the interaction of Step by Exposure Condition. However, the generalized linear mixed effects model analysis revealed no significant general effect of Exposure Condition, nor were the interactions between Step and Exposure Condition significant.

Table 3.6: Fixed-effect estimates of the performance of the non-native listeners in the noise listening condition in the phonetic categorization task

Fixed effect	β	SE	p
Intercept	-4.113	0.488	$<.001$
Exposure Condition	0.812	0.656	0.216
Step 2	1.018	0.399	0.011
Step 3	4.503	0.417	$<.001$
Step 4	5.277	0.432	$<.001$
Step 5	8.257	0.550	$<.001$
Exposure Condition x Step 2	0.690	0.523	0.187
Exposure Condition x Step 3	-0.464	0.537	0.387
Exposure Condition x Step 4	-0.961	0.550	0.081
Exposure Condition x Step 5	-1.070	0.733	0.144

In conclusion, lexically-guided perceptual learning in non-native listening plainly occurred when listeners were exposed to the target sound in clean listening conditions. The presence of intermittent noise weakened or inhibited lexically-guided perceptual learning to the extent that it either did not occur in the noise listening condition or, if present (Step 4 by Exposure Condition is marginally significant and the best-fitting model contains Exposure Condition and its interaction with Step), was restricted to only specific steps of the continuum.

3.3.5 Non-native listeners: comprehension

Since the ability to interpret the words containing an ambiguous sound is an important pre-requisite for lexically-guided perceptual learning to occur (e.g., Norris et al., 2003), listeners with a lower lexical proficiency or listeners who are worse at following the short story could have adapted to the ambiguous sound to a lesser extent. Potentially, this could have caused the differences between the noise and clean listening conditions for the non-native listeners.

The LexTALE scores were used as a measure of lexical proficiency and compared for the two non-native listener exposure groups. The final score on the LexTALE test was the percentage of correct responses corrected for the unequal proportion of words and nonwords in the test (Lemhöfer & Broersma, 2012).The average LexTALE score for the group of non-native listeners exposed to the short story in clean was 71.2 (SD=14.9), while the average score for the group of non-native listeners exposed to the short story in noise was 68.2 (SD=16.3). Both scores fall within the 60%-80% range, corresponding to the B2 or upper-intermediate level of English according to the Common European Framework of Reference (Lemhöfer & Broersma, 2012). The difference in LexTALE scores between the two non-native listener exposure groups was not significant: $t(76.657)=0.846$, $p=0.400$.

The proportions of the correct answers on the comprehension questions were calculated for the non-native listeners for the two listening conditions. The listeners in both listening conditions answered on average more than half of the comprehension questions correctly (clean listening condition: M=3.4 (SD=1.0); noise listening condition: M=3.1 (SD=1.1)). The difference in comprehension between the non-native listeners in the clean and intermittent noise conditions was not significant ($t(75.79)=1.42$, $p = 0.159$).

To further probe the role of proficiency in English and comprehension of the short story on the emergence of the lexically-guided perceptual learning effect, a new variable "the number of learning-consistent responses" was created (following Scharenborg et al., 2015). Learning-consistent responses are responses given in accordance with the exposure condition: /ɹ/ responses for the participants exposed to the /ɹ/-ambiguous version of the story, /l/ responses for the participants exposed to the /l/-ambiguous version of the story. In a subsequent mixed effects logistic regression analysis, the number of learning-consistent re-

sponses was used as the dependent variable, while proportion of correctly answered comprehension questions and lexical proficiency of the listeners (both scaled and centralized) were added as fixed factors. Subject was included as a random factor. Neither lexical proficiency (β=-0.048, SE=0.081, p=0.573) nor comprehension scores (β=0.117, SE=0.084, p=0.164) were shown to significantly influence lexically-guided perceptual learning. The difference in the emergence of the lexically-guided perceptual learning effect between the two different listening conditions for the non-native listeners could thus not be explained by a lower lexical proficiency or lower comprehension scores in noise compared to the clean listening condition.

3.4 Discussion and conclusions

The present study investigated the effect of intermittent noise on lexically-guided perceptual learning in native and non-native listening. We hypothesized that intermittent noise has a debilitating effect on lexically-guided perceptual learning, especially in the case of non-native listeners, due to the differences in the competition process in clean and noise for native and non-native listeners.

Retuning in clean listening conditions was demonstrated for both native and non-native listening in line with the results of numerous earlier studies showing a lexically-guided perceptual learning effect (e.g., for native listeners: Eisner & McQueen, 2006; Norris et al., 2003; Scharenborg et al., 2015; and for non-native listeners: Drozdova et al., 2016; Reinisch et al., 2013). Note that the here-presented new set of non-native data confirms those reported in Drozdova et al. (2016). Both the study by Drozdova et al. (2016) and the present study demonstrate that despite differences in native and non-native listening, relatively proficient non-native listeners are able to retune their non-native phonetic categories as a result of exposure to an ambiguous sound in a non-native language.

Native listeners in the present study also showed lexically-guided perceptual learning when background noise was present intermittently. Only one other study, to our knowledge, investigated the effect of the presence of background noise on the emergence of lexically-guided perceptual learning in native listening. Zhang and Samuel (2014) found that learning was blocked in the presence of noise during native listening, whereas we found the opposite. There is however an important differ-

ence between these two studies. During the exposure phase in the Zhang and Samuel study, the entire stimulus was masked by noise with the exception of the critical ambiguous sound. In our study, noise was far less prevalent, since it was never present on the words containing the ambiguous sound and most of the time also not on the words directly preceding and following the critical word. As Zhang and Samuel argued, the wide-spread presence of noise in the exposure increased the variability of the speech signal overall. Consequently, the variability of the ambiguous sound, which normally would trigger lexically-guided perceptual learning, could no longer be interpreted as a reliable cue to trigger retuning. In our study, the presence of noise might have increased the variability of the speech signal locally, but it did not reduce the reliability of the variability of the ambiguous sound as a cue to lexically-guided perceptual learning as evidenced by the fact that the native listeners still showed retuning in the intermittent noise listening condition.

Contrary to the native listeners, lexically-guided perceptual learning was affected and even largely inhibited for the non-native listeners when noise was present in the speech signal. While non-native listeners in the clean listening condition showed learning on all continuum steps (compared to only Steps 1 and 2 in the (different) non-native listener group in Drozdova et al., 2016, thus arguably showing a larger range of the effect than in the previous study), a difference between the /l/- and /ɹ/-exposure groups (non-significant however) was only visible on the first continuum steps for the group exposed to the short story in noise (see Figure 3.2). The statistical analysis did not show any significant effect of Exposure Condition (a replication of the findings of Drozdova, van Hout, & Scharenborg, 2015, on a different group of non-native listeners). On the other hand, as the best-fitting model contained the effect of Exposure Condition and its interaction with Step, we have to accept that incipient traces of Exposure Condition effects might be active. In short, non-native listeners, but not native listeners, demonstrate differences in lexical retuning between the clean and intermittent noise listening condition.

What causes this difference in the effect of the presence of intermittent background noise on lexical retuning in native and non-native listening? Speech recognition requires cognitive effort, while resources are in limited supply (Kahneman, 1973), with greater effort for adverse listening conditions than for optimal listening conditions. Potentially, the observed native versus non-native difference can be explained by a

greater difficulty, and consequently a larger cognitive effort, for the non-native listeners to recognize and integrate the words to understand the story in the exposure phase compared to the native listeners in the noise listening conditions. However, if that had been the case, comprehension of the short story in noise would have been worse than comprehension in the clean as more processing resources would have been involved in noise. This is not what was observed in the present study. Moreover, comprehension scores were shown to have no influence on the magnitude of lexically-guided perceptual learning, nor did lexical proficiency of the listeners, suggesting that the lack of lexically-guided perceptual learning in the noise listening condition was not caused by poor(er) understanding of the short story.

The observed native-non-native difference could have potentially been caused by different strategies applied by native and non-native listeners when listening occurred in noisy listening conditions. Mattys and colleagues (2010, 2011) found that when listening becomes harder, listeners seem to rely more on their strongest available cue: lexical knowledge for native listeners and acoustic detail for non-native listeners. This increased reliance on acoustic detail for non-native listeners could then inhibit lexically-guided perceptual learning. However, the non-native listeners in the Mattys, Carroll, Li, and Chan (2010) study also showed a larger reliance on acoustic detail compared to the native listeners in the clean condition, albeit to a lesser extent. One would therefore also expect a lack of retuning in the clean condition for the non-native listeners in the current study, but this is not what we observed.

Another explanation of the native-non-native difference between the noise and clean conditions is that it is simply a non-nativeness effect. As discussed in the Introduction, several differences exist between listening in a native and a non-native language. If the inhibition of the lexical retuning process for the non-native listeners in the intermittent noise condition were solely due to these differences then we would expect perceptual learning to be inhibited for the non-native listeners in the clean condition as well. This is not what we (and others, e.g., Reinisch et al., 2013) have found. The observed native-non-native distinction also cannot be merely explained by the presence of intermittent background noise. If the presence of noise was the overall cause, perceptual learning in the noise condition would have been blocked for the native listeners as well. This is not what we observed.

The absence of perceptual learning in non-native listening in the

presence of noise can possibly be explained by a larger effect of the presence of noise on the lexical competition process in non-native listening than in native listening (see also Scharenborg et al., 2017). Although the debilitating effect of noise on the competition process might be similar for native and non-native listeners (the current set-up of the experiment does not allow us to investigate this question), the consequences are larger for the non-native listeners. Not only are more candidate words considered for recognition in non-native listening compared to native listening (e.g., Broersma, 2012; Scharenborg et al., 2017; Weber & Cutler, 2004), the presence of noise has also been shown to increase the number of spuriously activated words to a larger extent in non-native listening than in native listening (Scharenborg et al., 2017). Jesse and McQueen (2011) found that information to disambiguate an ambiguous sound needs to be available timely, i.e., they found no retuning when the ambiguous sound was at the start of a word with the disambiguating lexical information only becoming available to the listener after hearing the ambiguous sound. Keeping multiple word candidates in memory slows down recognition of the word (Norris et al., 1995), including the word with the ambiguous sound in it. Consequently, the disambiguating lexical information becomes available later as well. Although the present study did not directly measure the time course of the activation of candidate words, the competition process in non-native listening in noise might have slowed down more than that in native listeners, due to the increase in the number of activated words in non-native listening compared to native listening in noise, to the extent that the crucial lexical information becomes available too late for lexically-guided perceptual learning to occur.

The difference in perceptual learning between the clean and noise listening conditions for the non-native listeners and the lack of a difference between the listening conditions for the native listeners were observed even though the ambiguous sounds used in the exposure and test phases were chosen on the basis of a pre-test with non-native listeners. This suggests that although a larger retuning effect might have been observed for the native listeners if the pretest had been carried out using native listeners (which consequently might have increased the observed native versus non-native difference), the chosen ambiguous sounds were ambiguous enough for the native listeners to induce lexically-guided perceptual learning in both the clean and the noise condition, for the *collect-correct* pair. We chose the most ambiguous steps on the basis of a pre-test with

non-native listeners to ensure that the chosen steps were indeed ambiguous for the non-native listeners, the group we were primarily interested in. As discussed in Drozdova et al. (2016), no lexically-guided perceptual learning was observed for the *alive-arrive* pair in the clean listening conditions for the native listeners, therefore this word pair was not further analyzed in the present article. Since the most ambiguous steps were chosen for each word separately on the basis of a pretest with non-native listeners, it is possible that the steps for *alive-arrive* were not well positioned for the native listeners. Post-hoc acoustic analyses (in Drozdova et al., 2016) indeed revealed that the first step of the *alive-arrive* continuum was more /ɹ/-like than the first step of the *collect-correct* continuum.

The present study demonstrates that noise, when present on parts of the speech stream, impedes lexically-guided perceptual learning for non-native to a larger extent than for native listeners. This native-non-native difference is argued to be due to the interplay of the effects of intermittent noise and non-native listening on the lexical competition processes during spoken-word recognition. We argue that the inhibition of phonetic category retuning in non-native listening in noise is due to the increase in number of activated words and the timing of the recognition of the word carrying the ambiguous sound in non-native listening compared to native listening in noise. When intermittent noise slows down the recognition of the critical word in non-native listening to the extent that the necessary lexical information to disambiguate the ambiguous sound arrives too late, phonetic category retuning is impeded.

CHAPTER 4

Processing and adaptation to ambiguous sounds during the course of perceptual learning

This Chapter has been adapted from
Drozdova, P., van Hout, R., & Scharenborg, O. (2016). Processing and adaptation to ambiguous sounds during the course of perceptual learning. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association (Interspeech 2016)*, 2811-2815.

Abstract

Listeners use their lexical knowledge to interpret ambiguous sounds, and retune their phonetic categories to include this ambiguous sound. Although there is ample evidence for lexically-guided retuning, the adaptation process is not fully understood. Using a lexical decision task with an embedded auditory semantic priming task, the present study investigates whether words containing an ambiguous sound are processed in the same way as "natural" words and whether adaptation to the ambiguous sound tends to equalize the processing of "ambiguous" and natural words. Analyses of the yes/no responses and reaction times to "natural" and "ambiguous" words showed that words containing an ambiguous sound were accepted as words less often and were processed more slowly than the same words without ambiguity. The difference in acceptance disappeared after exposure to approximately 15 ambiguous items. Interestingly, lower acceptance rates and slower processing did not have an effect on the processing of semantic information of the following word. However, lower acceptance rates of ambiguous primes predict slower reaction times of these primes, suggesting an important role of stimulus-specific characteristics in triggering lexically-guided perceptual learning.

4.1 Introduction

Previous research has demonstrated the ability of the human perceptual system to quickly adapt to ambiguously sounding items (Samuel & Kraljic, 2009). Norris et al. (2003) were the first to show that listeners use their lexical knowledge to interpret ambiguous sounds, e.g., an ambiguous final sound between /f/ and /s/ in *gira[f/s]* will be interpreted as an /f/ since *giraffe* is an existing English word and *giras* is not, while the same ambiguous sound in *bo[f/s]* will be interpreted as an /s/, since *boss* is an existing word and *bof* is not (Norris et al., 2003). Listeners adjust their phonetic category boundaries to include this ambiguous sound in their sound system (Clarke-Davidson et al., 2008). This mechanism is referred to as lexically-guided perceptual learning and is argued to aid listeners in adapting to unfamiliar speakers and accents (Norris et al., 2003; Reinisch & Holt, 2014).

Lexically-guided perceptual learning has been demonstrated using an exposure-test paradigm. In the exposure phase, participants listen to the ambiguous items, e.g., while performing a lexical decision task (Norris et al., 2003), and typically demonstrate learning in a subsequent phonetic categorization task. For lexically-guided perceptual learning to occur, ambiguous sounds should be embedded in real words (Norris et al., 2003) or phonotactically legal sequences (Cutler et al., 2008). Moreover, it has been shown that listeners, who accept more ambiguous items as real words, show a stronger learning effect (Scharenborg & Janse, 2013). This suggests that items with ambiguous sounds should be perceived as real words for learning to occur.

Although only a few studies have specifically looked at the time course of lexically-guided perceptual learning, it has been shown to be fast: exposure to as few as ten ambiguous items yields a stable learning effect (Kraljic & Samuel, 2007; Poellmann et al., 2011). Learning seems to occur in a step-wise manner: after exposure to ten items retuning did not get stronger with more exposure (Poellmann et al., 2011). The process of lexically-guided perceptual learning was further investigated by Scharenborg and Janse (2013) who showed that listeners increase their acceptance of words with an ambiguous sound as real words during the course of the exposure. The present study investigates in how far items containing ambiguous sounds are indeed perceived and processed as real, natural words. We do so by looking at the time-course of accepting words containing an ambiguous sound as a word, and by investigating

the spreading of activation to semantically-related words by words containing an ambiguous sound.

We use an auditory semantic priming paradigm within a standard lexical decision task as the exposure phase of a lexically-guided perceptual learning study. Multiple studies (e.g., A. M. Collins & Loftus, 1975) have demonstrated that processing of a word (target) is facilitated when it is preceded by a semantically-related prime. Primes in the present experiment contained an ambiguous sound [f/s], which either replaced all /s/ sounds while the /f/ sounds remained unchanged or replaced all /f/ sounds while all /s/ sounds remained unchanged. This set-up allowed us to compare reaction times and hit rates of words with ambiguous and natural sounds to study the recognition and (semantic) processing of "ambiguous" words in comparison to that of natural words. As mentioned by Andruski, Blumstein, and Burton (1994), studying the effect of the mismatch at the acoustic level in primes on the processing of the semantically-related targets can demonstrate differential activation of these words within the lexicon itself.

Since substitution of only one sound in words, or a mismatch in phonetic detail, hampers word processing (see McQueen, 2007 for an overview), we predict that words containing an ambiguous sound are accepted less often as real words and responded to more slowly than the same words with natural sounds. Moreover, we expect to find the same pattern of difference for the semantically-related target words (directly following the primes) due to a reduced semantic spreading by the ambiguous prime words. Additionally following Scharenborg and Janse (2013), we predict that listeners demonstrating more "natural-like" processing of ambiguous words exhibit more learning.

In order to investigate the time-course of accepting ambiguous items as real and natural words, we compare the difference in processing speed and recognition accuracy between words containing ambiguous and natural sounds in different parts of the exposure phase. We hypothesize that processing and recognition of the manipulated items will become more like the processing and recognition of their non-manipulated counterparts by the end of the exposure.

4.2 Method

4.2.1 Participants

Forty seven native Dutch participants (10 males, M_{age}=20.9, SD=2.0), recruited from the Radboud University Nijmegen subject pool, took part in the main experiment. Additionally, 11 native Dutch listeners (2 males, M_{age}=20.5, SD=0.5) participated in the pilot test of the stimuli. None of the pilot test participants took part in the main experiment.

4.2.2 Materials

For the exposure phase, 40 semantically related word-pairs were chosen from the Dutch Word Association Database (De Deyne & Storms, 2008). Crucially, the prime word of the pair contained either a word-final /f/ (e.g., *bankroof* (bank robbery); 20 words) or a word-final /s/ sound (e.g., *paleis* (palace); 20 words), while the target member of the word pair was semantically (highly) related to it (e.g., *geld* (money) and *koning* (king) for *bankroof* and *paleis*, respectively). Apart from the primes, no other words in the stimulus list contained /s/ or /f/. The lists of 20 /f/ and 20 /s/ prime words contained 8 one-syllable words, 8 two-syllable words, and 4 three-syllable words each. The same distribution was used for the target items. The chosen pairs were based on the "cue lookup" search mode in De Deyne and Storms (2008), which shows the ten most frequently generated associations for the cue word as well as the strength of the association. We used the highest associated word from the ten options which satisfied our constraints (word-final /s/ or /f/ in primes but none in targets, similar distribution of number of syllables per word, and the semantic association) as the target word. Due to the restrictions on the prime and target words, it was not possible to find all stimuli in the database. Another four word-pairs fitting the criteria were created and added to the stimulus set. The prime-target pairs used in the experiment are presented in Appendix D.

In addition, 60 Dutch words and 140 non-words were selected as fillers. The distribution of syllables was matched in both the critical items and fillers (i.e., 40% mono-, 40% bi- and 20% trisyllabic words). We divided the total number of stimuli into 14 blocks, each containing 20 items: three prime-target word-pairs, four filler words and ten non-words, except for the last block which contained one target word-pair.

Each block contained more than twice as many filler items as critical items to hide the associative relations in the prime-target word-pairs. In the final set of stimuli, the targets immediately followed the primes. This set up is similar to the one used by Schmidt, Scharenborg, and Janse (2015). All the items were produced by a female native Dutch speaker in a sound-attenuated booth at 44 kHz. The same speaker also recorded four minimal word-pairs for the test phase of the experiment: *brief-bries* (letter-breeze), *graf-gras* (grave-grass), *leef-lees* (live-read), and *lof-los* (praise-loose). Additionally, in order to create the ambiguous sound [f/s], 12 isolated syllable-pairs containing /s/ or /f/ with a vowel context identical to the vowel contexts in the primes were recorded (e.g., *eef-ees*).

4.2.3 Creating ambiguous stimuli

To create ambiguous versions of the prime words, the /s/ or /f/ sounds were excised from each recorded syllable and zero-padded with 25 ms of silence using a PRAAT (Boersma & Weenink, 2009) script and subsequently morphed using STRAIGHT (Kawahara et al., 1999) in Matlab. As a result of the morphing, an 11-step [f-s] continuum was created for each prime word separately, where step 0 was the most /f/-like and step 11 was the most /s/-like. To reduce an /s/-bias in some of the continua, sounds from these continua were remorphed using the original /f/ and step 7 of the created continuum. The most ambiguous sound between /f/ and /s/ was chosen on the basis of a pilot test with 11 native Dutch listeners. For the pilot test, the ambiguous sounds were spliced back to both members of the syllable-pairs (to avoid bias towards the /f/ or /s/ interpretation of the syllable). The pilot test included 240 items (five steps of each continuum presented four times). Items were presented to the participants binaurally through headphones in a sound-proof booth. Participants' task was to indicate whether the presented item contained an /f/ or an /s/ sound and press the corresponding button on the button box. The most ambiguous step was the step that received approximately 50% of /s/ and /f/ responses. This step of the sound was then spliced back into the prime words and used in the exposure phase in the main experiment. For the words in the test phase of the experiment, five versions were created using the most ambiguous step and the two steps preceding and following it.

4.2.4 Procedure

Two experimental lists were created for the exposure phase: in one list all primes with an /s/ sound were natural and all primes with an /f/ sound contained an ambiguous sound [f/s], while in the second list, all primes with an /s/ sound were ambiguous, while all primes with an /f/ sound were natural. The order of items in both lists was constant, and the same words served as primes in both lists. Primes that were ambiguous in one list were in their natural form in the other list, therefore providing a baseline for the comparison.

In the first part of the experiment, participants performed the lexical decision task. Stimuli were presented to the participants through headphones at a fixed mean intensity level of 70 dB. Listeners were instructed to react as fast as possible, and press the right button on a button-box if they thought the item they just heard was an existing Dutch word, and the left if they thought this word did not exist in Dutch. Set up of the exposure is presented on Figure 4.1.

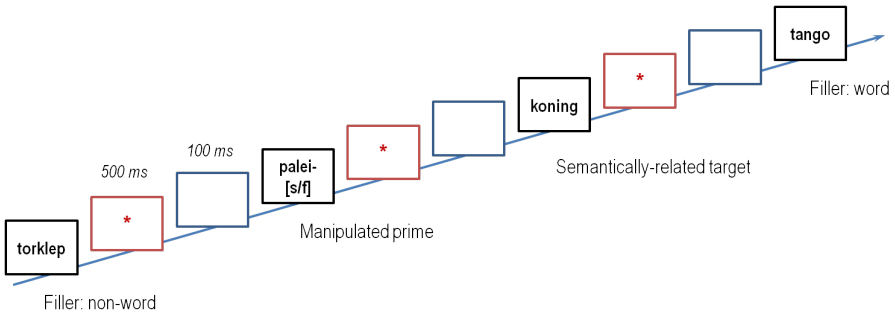

Figure 4.1: Set-up of the exposure phase of the experiment.

The subsequent phonetic categorization task consisted of 120 items, in which each ambiguous step of each minimal pair was presented 6 times. Listeners had to press the right button when hearing a word ending in an /s/-sound, and the left button if they heard a word ending in an /f/-sound. The /f/-interpretation of the minimal pair was shown on the left of the computer screen, and the /s/-interpretation of the minimal pair on the right side. The whole experiment took approximately 20 minutes.

4.3 Results

4.3.1 Phonetic categorization task

To investigate the processing of words with an ambiguous sound during lexically-guided perceptual learning, it is necessary to first establish whether lexically-guided perceptual learning occurred. Responses of the listeners in the phonetic-categorization task were analyzed using generalized linear mixed effect models (Baayen et al., 2008). The dependent variable was the number of /s/-responses. The analysis started with the model including Exposure Condition (whether participants were exposed to /s/-ambiguous or /f/-ambiguous tokens), Step on the /f/ to /s/ continuum (as a nominal variable) and their interaction as fixed predictors. Subject and Word were included as random factors.

Figure 4.2 shows the proportion of /s/ responses in the phonetic-categorization task for the five test steps. The responses of the participants exposed to the items where all /f/ sounds were ambiguous are plotted with the dotted line with squares, the responses of the other group with the solid line with circles. The difference between the two lines represents the lexically-guided perceptual learning effect.

As shown in Figure 4.2, participants exposed to the words with an ambiguous /s/-sound gave significantly more /s/ responses in the phonetic-categorization task than the participants exposed to the words with an ambiguous /f/-sound. This observation was confirmed by the statistical analysis which showed a significant interaction between Exposure Condition and Step 3 (β=0.783, SE=0.226, p <.001), Step 4 (β=0.784, SE=0.232, p <.001), and Step 5 (β=0.850, SE=0.251, p <.001). Note that the main effect of Exposure Condition was marginally significant (β=0.718, SE=0.430, p=0.095). These systematic differences between the two exposure groups indicate that the listeners showed lexically-guided perceptual learning, and thus that the ambiguous sound was included in the sound system of the listeners.

4.3.2 Lexical decision task

To investigate the extent to which items with an ambiguous sound are processed and recognized as real words, responses of the listeners to the primes and targets in the lexical decision task were analyzed. Recognition of the ambiguous primes was investigated by comparing the hit

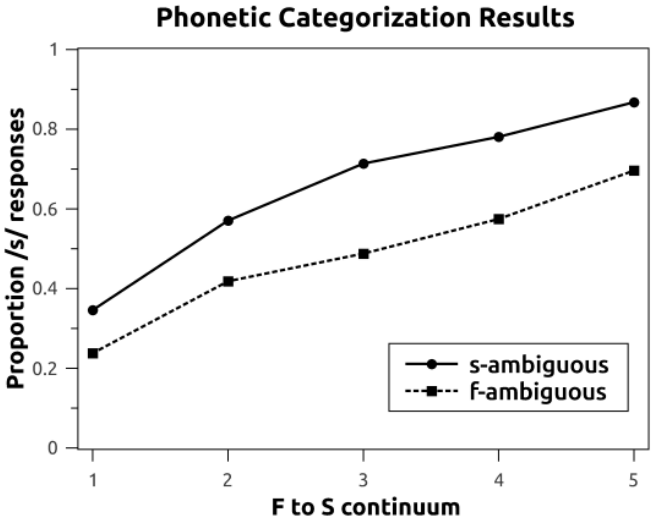

Figure 4.2: Proportion of /s/ responses of the two exposure conditions in the phonetic-categorization task.

rates (number of "yes" responses) and reaction times to the manipulated primes to those of the natural primes. Semantic spreading of activation of the ambiguous primes was investigated by comparing the hit rates and reaction times on the target items preceded by the manipulated and by the natural semantically-related prime. To investigate the time course of lexically-guided perceptual learning, the 40 prime-target pairs were subdivided into 4 equal-sized bins. Finally, following Scharenborg and Janse (2013), the number of hits was used as a predictor in a separate analysis to investigate whether more "natural"-like processing resulted in more learning.

Analyses of hit rates

One word-pair, *poos-tijd* (a while - time), was excluded from the analyses, since even the natural variant of the word *poos* was accepted as a real word less than 50% of the time. Table 4.1 provides the mean proportions of the "yes" responses and their standard deviations (in brackets) for the natural and manipulated primes and their semantically-related targets.

Table 4.1: Mean proportions of "yes" responses for different types of primes and targets.

Manipulation	Hits (primes)	Hits (targets)
Natural	0.96 (0.19)	0.996 (0.07)
Manipulated	0.91 (0.29)	0.993 (0.08)

Generalized linear mixed effect models (Baayen et al., 2008) were used to analyze the hit rates with either hit rates for primes or targets as a dependent variable. Manipulation of the prime (whether the prime contained an ambiguous sound), Bin Number and the interaction between them were used as fixed factors, and Subject and Item were added as random factors.

In the hit rate analysis for primes, only Manipulation was shown to be a significant predictor of the number of "yes" responses: participants accepted fewer manipulated than natural items as real words (β=-1.020, SE=0.225, p <0.001). Bin Number and its interaction with Manipulation did not reach significance. When Bin Number was added to the model as a nominal variable with Bin Number 4 on the intercept, there was a significant interaction between Manipulation and Bin Number 2 (β=-1.493, SE=0.617, p=0.016) and 3 (β=-1.543, SE=0.632, p=0.015). Although the interaction between Bin Number 1 and Manipulation did not reach significance (β=-1.073, SE=0.724, p=0.138), it was in the same direction. Thus, where no significant difference was observed between the manipulated and natural primes in Bin Number 4, there were significant differences in acceptance rates for primes with the ambiguous sound and the same primes with the natural sound in the earlier bins. Recognition of the manipulated prime thus became more natural in the last Bin (i.e., with the last 10 items).

In the hit rates analysis for the targets, no fixed factors reached significance. Throughout the lexical decision task, recognition of the target words was high, irrespective of the type of the prime. A final analysis with the number of learning-consistent responses, i.e., responses given in accordance with the exposure condition (e.g. /s/ responses for the participants exposed to the /s/-ambiguous list), as a dependent variable and Acceptance Rate (i.e., the proportion of hits for the manipulated version of each prime) as fixed factor showed that a higher acceptance rate of

ambiguous words as real words leads to a larger marginally significant learning effect (β=3.480, SE=1.812, p=0.055).

Analyses of reaction times

Table 4.2 provides the average reaction times for primes and targets with their standard deviations. All reaction times deviating more than two standard deviations from the mean were excluded, and only reaction times of the primes and targets which were accepted as real words were analyzed. Log transformed reaction times for either prime or targets were used as a dependent variable in the analyses. In addition to the factors mentioned in Section 4.3.2.1., the Acceptance Rate and its interaction with Manipulation were included in the analysis to investigate whether ambiguous primes which were less easily accepted as real words also exhibited larger differences in reaction times between their manipulated and non-manipulated versions and less spreading of activation to the target words.

In the reaction times analysis of the primes, again a main effect of Manipulation (β=0.066, SE=0.007, t=9.7) was observed. Primes with manipulated sounds were reacted to more slowly than primes with natural sounds. Moreover, the interaction between Acceptance Rate and Manipulation reached significance (β=-0.287, SE=0.098, t=-2.9): lower acceptance rates of ambiguous primes related to larger differences in reaction times from their natural counterparts. The difference in reaction times between primes with an ambiguous and natural sound did not change during the task, the factor Bin Number and its interaction with Manipulation did not reach significance.

Similar to the hit rate analysis, none of the factors in the reaction time analysis of the targets reached significance. There were no differences in the speed of processing of the targets, irrespective of whether they were preceded by natural or manipulated primes. Moreover, processing of the target words preceded by ambiguous and natural primes remained similar throughout the task. Finally, reaction time differences between the manipulated and natural versions of the words did not predict perceptual learning.

Table 4.2: Mean reaction times for different types of primes and targets.

Manipulation	Hits (primes)	Hits (targets)
Natural	954.35 (178.93)	836.18 (180.72)
Manipulated	1016.18 (186.41)	832.15 (183.87)

4.4 General discussion and conclusions

The present study investigated the perception and processing of words containing ambiguous sounds during the course of lexically-guided perceptual learning using an auditory semantic priming paradigm. Two questions were studied: whether words with an ambiguous sound are recognized and processed in the same way as their non-ambiguous counterparts, and what the time course is of the adaptation to an ambiguous sound. Our hypothesis that primes containing an ambiguous sound will be accepted less often as real words than their natural counterparts and will be reacted to slower was confirmed. The manipulation of a sound in a word led to an increase in reaction times and fewer "yes" responses. This finding is in line with existing literature showing that changes in phonetic details of phonemes interfere with word recognition (e.g., McQueen, 2007; Whalen, 1991).

Despite the differences in the processing of ambiguous primes compared to natural primes, there were no differences in acceptance rates or reaction times between target words preceded by ambiguous primes and targets preceded by natural primes. This suggests that although the manipulated primes were perceived as less natural and processed slower than the natural primes, this did not have an effect on the processing of semantic information. Possibly, the presence of an ambiguous sound slows down the build-up of the activation of the word so that the threshold for word recognition is reached later. The build-up is, however, fast enough so that activation can spread to semantically-related words. To specifically tap into the priming effect, the priming effect of ambiguous and natural primes could be further investigated by including a set of word pairs, where the target items are preceded by a non-related word (similar to Andruski et al., 1994).

We hypothesized that processing and recognition of primes with an ambiguous sound would differ from that of natural primes at the start of

the exposure phase and would become more like the processing of natural words towards the end of the exposure phase. The results showed that although processing of ambiguous words remained slower than that of natural words, recognition did become more natural-like. Similar to Scharenborg and Janse (2013), listeners increase their acceptance of words with an ambiguous sound as real words during the course of the exposure phase. Participants' recognition of the prime containing the ambiguous sound changed to natural-like after exposure to approximately 15 items. This is in line with Kraljic and Samuel (2005) and Poellmann et al. (2011) who observed learning after exposure to 10 items. Moreover, this learning seemed to occur in a step-wise manner, like was found by Poellmann et al. (2011), as no convergence in the hit rates of the natural and ambiguous words was observed in the bins prior to the final bin. In line with Scharenborg and Janse (2013), listeners who accepted more ambiguous words as real words showed a larger learning effect.

Ambiguous primes with lower acceptance rates were found to yield relatively longer processing times than their natural counterparts, as shown by the significant interaction between Manipulation and acceptance of the ambiguous version of the prime in the reaction time analysis. Some ambiguous items were thus more difficult to process and recognize than other ambiguous items, despite being manipulated in a similar way. Supposing that only ambiguous stimuli recognized as real words induce retuning, this is an important finding suggesting that stimulus-specific characteristics (e.g., the size of the lexical neighborhood) may influence lexically-guided perceptual learning. This factor of word characteristics adds to a growing list of factors known to influence lexically-guided perceptual learning, including listener-related factors (e.g., listeners' acceptance of an ambiguous item as a word (Scharenborg & Janse, 2013), attention-switching control (Scharenborg et al., 2015), or environment-induced factors (e.g., pen in the mouth of the speaker (Kraljic, Samuel, & Brennan, 2008)), and noise (Drozdova et al., 2015; Zhang & Samuel, 2014)).

In conclusion, there are clear differences between the processing and recognition of words containing an ambiguous sound and the same words with a natural sound. The slower and less accurate processing of ambiguous words, however, does not interfere with semantic processing of the ambiguous words. Moreover, adaptation to the ambiguous sounds is (again) fast and quickly results in a recognition process that is similar to that of natural words.

CHAPTER 5

L2 voice recognition: the role of speaker-, listener- and stimulus-related factors

This Chapter is based on
Drozdova, P., van Hout, R., & Scharenborg, O. (2017). L2 voice recognition: the role of speaker-, listener- and stimulus-related factors. Manuscript accepted for publication in *the Journal of the Acoustical Society of America*.

Abstract

Previous studies examined various factors influencing voice recognition and learning, and they have obtained mixed results. The present study investigates the separate and combined contribution of these various speaker-, stimulus-, and listener-related factors to voice recognition.

Dutch listeners, with arguably incomplete phonological and lexical knowledge in the target language, English, learned to recognize the voice of four native English speakers during four-day training. The training was successful and listeners' accuracy was shown to be influenced by the acoustic characteristics of speakers and the sound composition of the words used in the training, but not by lexical frequency of the words, nor the lexical knowledge of the listeners or their phonological aptitude. Although not conclusive, listeners with a lower working memory capacity seemed to be slower in learning voices than listeners with a higher working memory capacity. The results reveal that speaker-related, listener-related, and stimulus-related factors accumulate in voice recognition, while lexical information turns out not to play a role in successful voice learning and recognition. This implies that voice recognition operates at the prelexical processing level.

5.1 Introduction

Recognizing voices is a prodigious human cognitive ability. Recognition of the mother's voice in infancy has a key role in children's emotional, social, and cognitive functioning (Abrams et al., 2016). The ability of adults to recognize people by voice forms a crucial social skill (Perrachione et al., 2011), and, in general, contributes to the perception of interlocutors' emotional states and personal identities (Nygaard, 2005; Sidtis & Kreiman, 2012). Moreover, voice familiarity has been shown to facilitate word recognition and processing (Nygaard & Pisoni, 1998; Nygaard et al., 1994).

While the importance of the ability to recognize voices is beyond dispute, the factors influencing successful voice recognition are still unclear. Several types of factors have been investigated, but the obtained results were mixed. The present study groups these various factors into three categories: speaker-related, listener-related and stimulus-related, and investigates the role of these three groups of factors on speaker recognition by second language (L2) listeners, in order to shed light on their separate and combined contribution to successful voice recognition.

The first group of factors which have been shown to influence voice recognition is related to the acoustic characteristics of speakers' voices. Laver (1968) distinguished three types of information conveyed in a speaker's voice: biological (gender, age, size), psychological (emotional state), and social information (regional origin, social group, profession). This information can be expressed in diverse voice quality features (e.g., loudness, pitch, phonation types, nasalization). These features are used to a different extent in voice recognition and differentiation (see Baumann & Belin, 2010 for an overview), with fundamental frequency (F0) being the most prominent one for distinguishing voices, while the importance of other characteristics such as frequencies of the main formants (F1, F2, F3), jitter, and shimmer are dependent on the type of speaker (e.g., male or female, pathological or normal voices). In this study, we specifically investigate the contribution of two speaker-related factors: fundamental frequency (average, minimum and maximum) and average word length.

Interestingly, speakers were shown to vary in their identifiability which depended not only on the quality of their voices, but also on which specific speech sounds were produced (Amino & Arai, 2008; Andics, 2013; Andics, McQueen, & Van Turennout, 2007; Bricker & Pruzansky, 1966).

Not all sounds are equally effective in conveying speaker-specific information. Research in both automatic speech recognition (Eatock & Mason, 1994; Gallardo, Möller, & Wagner, 2015) and human speech recognition (Amino & Arai, 2008; Amino, Sugawara, & Arai, 2006) provide a ranking of sounds contributing to talker-identification, showing that nasal consonants and vowels are more informative than other sounds for the human identification of speakers. Vowels outperform consonants due to their combination of fundamental frequency (F0) and rich harmonic structure (Owren & Cardillo, 2006), while nasals outperform other consonants due to the speaker-specific characteristics of the resonating shapes involved and the timing of the velum movements for the production of nasals which is consistent within a speaker and differs among speakers (Amino, Arai, & Sugawara, 2007). The number of nasals and vowels in the word was investigated as one of the stimulus-related factors.

Most stimulus-related and listener-related factors are however not so easy to disentangle: stimulus-related factors relate to the actual linguistic information available in the stimuli, whereas listener-related factors relate to what extent listeners differ in their ability to use this information. Voice learning studies demonstrated that listeners can learn to recognize talkers without any access to linguistic content, e.g., in time-reversed speech (Bricker & Pruzansky, 1966; Sheffert et al., 2002) or in a completely unfamiliar language (Winters et al., 2008). At the same time, speaker-specific (also referred to as indexical information; Abercombie, 1967) and linguistic information in the signal are not completely independent. On the one hand, being familiar with a speaker's voice facilitates word recognition (Bradlow, Nygaard, & Pisoni, 1999; Levi et al., 2011; Nygaard & Pisoni, 1998), on the other, linguistic knowledge can support processing of indexical information as well (Goggin et al., 1991; Bregman & Creel, 2014; Winters et al., 2008).

Goggin and colleagues (Goggin et al., 1991) showed that voice recognition is more accurate when listeners understand the language being spoken: monolingual English listeners identified bilingual English-German speakers better when they spoke English than when they spoke German, while the reverse pattern was true for monolingual German listeners. At the same time, Bregman and Creel (2014) demonstrated that Korean-English bilinguals are faster in learning to recognize voices speaking in their first (Korean) than in their second language (English), and that the rate of learning in recognizing voices talking in their second language is modulated by their age of acquisition. Winters and colleagues

(Winters et al., 2008) studied whether native English listeners trained to recognize speakers either with German or English stimuli could generalize their knowledge and correctly identify the same speakers when they spoke the language the listeners had not been trained on (English or German respectively). While native English listeners could identify the same listeners significantly better when they spoke in English than when they spoke in German, no differences were observed for listeners trained in German. The authors concluded that listeners made use of language-dependent indexical cues to identify speakers speaking a familiar language. These studies show that although there is enough language-independent information in the signal to successfully recognize speakers when the signal lacks linguistic information, indexical information to recognize voices is not language independent. Voice recognition performance is better when listeners are (more) familiar with the language being spoken.

The advantage of a familiar language in voice recognition can possibly be explained by listeners' understanding of what is being said, or, in other words, access to lexical information. If so, lexical frequency of words and lexical knowledge of listeners could potentially influence voice recognition. School children indeed have been found to show improved sensitivity to talker-differences in highly familiar words (Levi & Schwartz, 2013). Moreover, school children showed a larger "voice familiarity" effect (i.e., better word recognition performance for familiar than unfamiliar voices) in highly familiar words than in less familiar words (Levi, 2015). Lexical frequency highly correlated with word familiarity ratings for those children. However, no effect of lexical frequency was found for voice learning and recognition by adult listeners (Winters et al., 2008) or children (Levi, 2015, 2014), suggesting that phonological rather than lexical knowledge plays the most important role in the perception of speaker information. This explanation was put forward by Perrachione and Wong (2007) who explained the better performance of their listeners in their native compared to an unfamiliar language by arguing that some degree of proficiency in the language is needed to gain access to inter-talker phonetic variability. Moreover, Perrachione and colleagues (Perrachione et al., 2011) observed impaired performance of dyslexic listeners in voice recognition tasks, which was correlated with their degree of phonological impairment. They concluded that dyslexic listeners fail to learn speaker-specific representations of phonetic consistency which leads to the observed voice recognition problems. Creel and Jimenez (2012) offered a similar explanation after finding differences in voice learning between

adults and pre-school children. They argued that pre-school children experience problems in encoding speaker information, since they are worse than adults in recognizing speech patterns and are still learning acoustic cues mapping to speakers' identity. In discussing larger voice familiarity effects in high-familiar words for children which was attributed to word frequency, Levi (2015) suggested that children might be less sensitive to acoustic-phonetic details in the input when they are exposed to low-familiar words, while the lack of the lexical frequency effects in adults might be connected to the fact that even low-frequent words are not unfamiliar enough. At the same time, knowledge of the phonological structure alone cannot fully explain the findings in voice recognition studies. Linguistic similarity between familiar and unfamiliar languages does not seem to modulate the language familiarity effect: Chinese listeners identified German speakers better than Spanish listeners did and they even outperformed English listeners whose native language is (far) more phonologically related to German than Chinese is (Köster & Schiller, 1997). In our experiment, lexical frequency of the word was used as a stimulus-related factor, whereas lexical proficiency in the L2 language was used as a listener-related factor.

Apart from listeners' lexical knowledge and language experience, a number of other listener-related factors have been shown to influence voice recognition performance. Since listeners have to learn to recognize previously unfamiliar speakers, working memory capacity can potentially influence the degree of voice learning and accuracy of voice recognition. Working memory is associated with the short-term storage of incoming information and its manipulation (Levi, 2014). Previous studies (Bregman & Creel, 2014; Levi, 2014) indeed found a positive relation between the component of working memory termed Phonological Loop (Baddeley, 1986; Baddeley & Hitch, 1974) responsible for short-term storage of auditory information, with the speed of voice learning in bilingual listeners (Bregman & Creel, 2014) and accuracy of voice recognition in native school age children (Levi, 2014). At the same time, another component of working memory namely Central Executive, responsible for manipulating the upcoming information and divided attention and controlling the Phonological Loop, was found to negatively impact voice recognition performance of children on the last day of voice training (Levi, 2014). Furthermore, individuals' phonological memory and awareness have also been found to play a role in a voice recognition. This pattern was observed for both dyslexic (Perrachione et al., 2011) and non-dyslexic lis-

teners (Jimenez, 2012). Perrachione and colleagues showed that voice recognition performance of dyslexic listeners correlated with their results on the Comprehensive Test of Phonological Processing (Torgesen, Rashotte, & Wagner, 1999), which measures phonological awareness and phonological memory. A similar correlation of phonological processing and the ability to recognize voices was demonstrated in an experiment with non-dyslexic listeners (Jimenez, 2012). The contributions of working memory capacity and phonological aptitude were investigated as listener-related factors in the present study.

Given the observed role of phonological knowledge and memory and the mixed findings about the role of lexical knowledge in voice learning, recognition and discrimination, it is surprising that there is a lack of voice-learning studies with L2 listeners. Testing this group of listeners with their incomplete L2 lexical and phonological knowledge in comparison to native listeners of the language can provide new insights into the role of the factors influencing the encoding of voice information. L2 listeners are less familiar with the sound and lexical structure of their second than their first language. Moreover, languages differ in their, partly non-linguistic, acoustic parameters, which could be used for voice recognition (Johnson, Westrek, Nazzi, & Cutler, 2011). For instance, F0 has an overall wider range in English than in Dutch (Chen, Gussenhoven, & Rietveld, 2004; B. Collins & Mees, 1999). Bregman and Creel (2014) hypothesized that in order to discriminate talkers speaking in a particular language, listeners need to be familiar with talker-varying characteristics unique to that language. It is, therefore, possible that L2 listeners have some difficulty using acoustic cues to identify voices in a non-native language. Using L2 listeners may allow us to look deeper into the role of proficiency and lexical knowledge in voice learning and recognition. As suggested in a recent study by White and colleagues (White, Yee, Blumstein, & Morgan, 2013), weak lexical representations result in the reduction of sensitivity to phonetic detail not only in children, but also in adults. If this is the case, L2 listeners would be less able to exploit acoustic-phonetic details in low-frequent than in high frequent words, resulting in lexical knowledge and lexical frequency effects in voice learning and recognition in the L2 language. There are only two voice learning study with L2 listeners that we are aware of. Perrachione and Wong (2007) showed that Mandarin listeners residing in the US, and speaking predominantly English in their daily life, recognized voices speaking in Mandarin better than voices speaking in English at the beginning of

a voice learning paradigm, but this difference disappeared in the latter sessions. Bregman and Creel (2014) found that the speed of learning to recognize a voice positively correlated with the age of acquisition of the second language. However, unlike in the current study, no direct measure of second language proficiency was used in these studies.

The aim of the present study is to investigate the role of speaker-, listener-, and stimulus-related factors in voice recognition and learning in L2 listeners. To that end, a group of Dutch participants was trained to recognize four previously unfamiliar voices speaking in British English during a four-day training period (similar to Nygaard & Pisoni, 1998, who used a 10-day training period). As Nygaard & Pisoni (1998) trained their participants to recognize the voices of ten speakers, and the current study included only four speakers, four days rather than ten days of training were used (see also Winters et al., 2008 for a successful 4-day voice learning training). The voice recognition accuracy and learning progress per day of the listeners was measured in relation to the speakers' voice characteristics (minimum, maximum, and average fundamental frequency, and average word length for each speaker) and stimulus characteristics (lexical frequency and the number of phonemes carrying indexical information). Moreover, participants' lexical knowledge was measured with the LexTALE test (Lemhöfer & Broersma, 2012), while their phonological aptitude was measured with the Llama-D test (Meara, 2005). The computerized variant of backward Digit Span (Wechsler, 2008) with visual presentation of the stimuli and written recall was used to assess the role of working memory capacity (namely the Central Executive component of working memory) following previous studies (e.g. Alloway, Gathercole, Willis, & Adams, 2004; Bull, Espy, & Wiebe, 2008; Levi, 2014; Rosenthal, Riccio, Gsanger, & Jarratt, 2006).

5.2 Method

5.2.1 Experimental set-up

The experiment contained four sessions divided over four consecutive days. Each session combined a training and a test phase. The procedure used in this experiment follows the methodology originally developed by Nygaard and Pisoni (1998) which is applied and adapted in later studies to investigate voice learning and voice familiarity effects (e.g., Levi, Winters, & Pisoni, 2008; Levi et al., 2011; Levi, 2014). Table 5.1

gives an overview of the tasks and the number of words included in each task on each day of the experiment.

Table 5.1: Experimental set-up on each training day. The number of words in each task is included in brackets.

Day	1	2	3	4
Tasks	Familiarization (24) Feedback (64) Test(64) Llama LexTALE	Feedback Test Digit Span	Feedback Test	Feedback Test
Duration (min)	45	30	30	30

5.2.2 Materials

76 mono- and 76 bisyllabic English nouns were chosen from the SUBTLEX UK database (Van Heuven, Mandera, Keuleers, & Brysbaert, 2014). Both bisyllabic and monosyllabic sets contained words of different frequencies (the distribution was relatively similar for both monosyllabic and bisyllabic words): from 1.02 per million for the lowest frequency word in the set (*sob*) to 589 per million for the highest frequency word in the set (*end*). All words were content words and were judged as familiar to the L2 participant's group by the authors.

The words were recorded by 12 native British male speakers, who at the time of the experiment, were living in or visiting the Netherlands. They came from different parts of Great Britain, and were between the ages of 21 and 33. Table 5.2 presents for each speaker the average, standard deviation and range (minimum and maximum) of the F0 as measured by Praat (Boersma & Weenink, 2009) in Hz, and the average word length. Each speaker read the word list aloud twice (second time in the opposite order). The speakers were recorded individually in a sound-proof booth with a Sennheiser ME 64 microphone at a sampling frequency of 44100 Hz. Words which were mispronounced or produced too quietly were recorded again. The words were then excised from the resulting audio files using a Matlab (The MathWorks Inc., 2013) script, and the segmentations were subsequently manually checked using Praat

(Boersma & Weenink, 2009). All speakers were rewarded 5 Euro for half an hour of recording time.

Table 5.2: Characteristics of the talkers used in the experiment.

	F0				Word Length (ms)	
Talker	Mean	SD	Minimum	Maximum	Mean	SD
1	107	15	89	137	585	116
2	98	16	73	129	556	122
3	153	27	115	198	490	106
4	119	13	104	143	527	118
5	90	26	73	142	516	103
6	137	19	116	162	579	137
7	114	20	94	144	437	97
8	148	15	133	174	526	110
9	156	23	103	230	569	141
10	122	21	97	155	624	124
11	115	20	93	145	431	96
12	142	11	126	165	548	119

In order to study the general process of voice learning, as well as to be able to include individual characteristics of voices in the analysis, the listeners were trained on different sets of speakers. This is in contrast to previous studies which provided all listeners with the same set of speakers (e.g., Nygaard & Pisoni, 1998; Winters et al., 2008). Listeners were trained to recognize four speakers from the set of 12 different speakers; however, Speaker 1 was the same for all participants. This was necessary for another experiment, which is not reported here. The other three speakers were chosen from the remaining 11 speakers, in different combinations. Eleven combinations (lists) were created (e.g., list 1: speaker 1, 5, 8, 9; list 2: speaker 1, 11, 7, 12, etc.). Speakers 2-12 occurred three times in all the lists in different positions.

Twenty-four words from the voice learning set were used on the first session in a familiarity phase (12 monosyllabic and 12 bisyllabic words). The remaining 128 words were semi-randomly divided over the stimuli for the feedback and test session for each day of training, so that both the feedback and test phases contained an equal number of bisyllabic

and monosyllabic words with comparable frequency. Following Nygaard and Pisoni (1998), the same words were used in the voice learning part of the experiment on each day, but the speaker producing the word varied per day (e.g., if the word is pronounced by Speaker 1 on Day 1, it will be pronounced by Speaker 2 on Day 2). Each listener heard each word of a particular speaker only once during the course of the experiment. Different from Nygaard and Pisoni (1998), the division of the words into the set used for the feedback and the test phases differed for each day to ensure generalization of learning. Moreover, two different orders of the stimuli presentation were used (i.e. the stimuli which were presented to half of the participants on the first day of the training, were presented to the other half of the participants on the fourth day of the training).

5.2.3 Participants

Forty-five (10 males, M_{age}=22.5, SD=2.4) native speakers of Dutch with no reported history of learning or hearing disorders were recruited from the Radboud University Nijmegen subject pool. All participants had a minimum of eight years of formal training in English and possessed a "VWO" (i.e., pre-university education) diploma, meaning a B2 or higher level of English according to the European Framework of Reference. Additionally, 16 participants (2 males, M_{age}=21.9, SD= 2.8) took part in the pre-test of the stimuli and experimental set-up. None of the participants who participated in the pre-test, took part in the main experiment. All participants received study points or 30 Euro for their participation.

5.2.4 Procedure

All participants were tested individually in a quiet sound-attenuated booth. The stimuli were presented to them binaurally through headphones. The intensity level of all the stimuli was set at 70 dB SPL. The experiment was administered with Presentation software (Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com).

Training

In the training phase on Day 1, participants were first familiarized with the speakers. They heard a sequence of five words produced by each of the four speakers (different words for different speakers) followed by one word from each of the speakers, with the name of the speaker appearing

on the screen. This procedure was repeated twice. The listeners' task was to memorize the name and the voice of the speaker. The task was self-paced and listeners had to press a button when they were ready for the next word. The familiarization phase only occurred once, on Day 1.

After the familiarization phase on Day 1 and at the start of Days 2-4, listeners had to complete the feedback phase of the task. In this phase, participants heard a word produced by one of the four speakers, and had to choose one name from the (earlier introduced) four names which were presented on the screen. If the choice was correct participants saw "correct" appearing on the screen; if a mistake was made, the correct name appeared on the screen. Participants were instructed to press one of the four buttons on the button box corresponding to the name of the speaker. They were told to react as quickly as possible, but at the same time to minimize mistakes. The position of the names on the screen changed each day to ensure deeper learning.

Test

The test phase was similar to the feedback phase but without any feedback on the answers. Participants listened to the word and again had to choose one name from the four names appearing on the screen and press the button on the button box corresponding to the name of the speaker they thought they just listened to. After the response was given, participants moved to the next word. Only the responses of participants from the test phase of the experiment were analyzed.

Cognitive tests

After completing the three voice learning tasks on Day 1, participants had to perform the Llama and LexTALE tests. At the end of the second day of the training, participants completed the backward Digit Span task.

Llama-D test is part of a battery of language aptitude tests developed by lognostics (Meara, 2005). The test measures the ability of the listeners to learn, recognize and discriminate phonological sequences (Meara, 2005; Granena & Long, 2013). In the Llama-D test participants listen to a set of ten (non)-words in an unfamiliar language. Their task is to listen to the words carefully since in the second part they have to decide, by pressing one of two buttons, whether the word they hear then was

already presented to them in the first part. Participants both gain points for correct responses and lose points if they make an error.

LexTALE (Lemhöfer & Broersma, 2012) is a visual unspeeded lexical decision task for advanced learners of English. Participants decide by pressing one of two buttons whether the word they see is an existing word in English. The test consists of 60 trials, designed to test vocabulary knowledge of medium to highly proficient L2 speakers of English.

In the backward Digit Span task (Wechsler, 2008), which measures working memory capacity, participants see a number of digits on a screen. The digit string increases with one digit every two trials, starting from two digits and ending with a sequence length of seven digits. Each sequence of digits is presented by consecutively showing the digits on the screen for one second with a one-second-interval before the next digit of the sequence is shown. The task of the participant is to memorize the sequence and type in the digits in reverse order.

Following Neger, Rietveld, and Janse (2014) each participant was presented with all sequence lengths irrespective of their performance on earlier trials. The visual form of digit presentation was chosen over an auditory presentation since the visual backward Digit Span task is considered optimal in multilingual settings as it allows one to tease apart non-native proficiency of the listeners and their working memory capacity (Owren & Cardillo, 2006).

5.3 Results

Due to missing data on one of the training days, data from five participants had to be excluded from the analysis. Data from the remaining 40 participants were analyzed with mixed effects logistic regression analysis (Jaeger, 2008) in R (version 3.3.2). The analysis was conducted in several steps. First, we want to establish whether the L2 listeners managed to learn to recognize previously unfamiliar voices during the training. Responses of the participants in the test phase of the experiment were coded as 1 if the response of the participant corresponded to the correct name of the speaker, and 0 if the answer was wrong. Accuracy (whether the response was correct or incorrect), thus, served as a dependent variable in the analysis. To assess the effect of the day of the training on the number of times the speaker was recognized correctly, Day (1, 2, 3, 4) was included as a fixed factor, while Subject, Word, List (the combination

of speakers that the participant listened to) and Speaker were included as random factors. Additionally, a by-Subject random slope for Day was introduced to account for differences over time in voice learning caused by differences across participants.

In the second step of the analysis, the effect of the various speaker-, stimulus-, and listener-related factors was investigated by adding the three types of factors group-wise to the previous best model. Each factor was added as a main factor and in interaction with Day to investigate the contribution of this factor to participants' voice recognition performance and the speed of voice learning. For the speaker-related factors, a measure of predictability of each voice based on its acoustic characteristics (average, minimum and maximum F0 and average word length) was included in the analysis. As stimulus-related characteristics, the number of phonemes in the word carrying indexical information (Amino & Arai, 2008; Eatock & Mason, 1994; Gallardo et al., 2015) and lexical frequency were added to the model. Finally, the results of the listeners on the language and cognitive tests were calculated and included in the statistical model as listener-related characteristics. The new model was compared to the previous best-fitting model to evaluate whether the new model explains significantly better the variation in the data than the previous model. After evaluating the model fit, the significance of the effects is established, following Snijders and Bosker (2012). All continuous variables were centered and scaled. All steps of the analysis are explained in more detail below.

5.3.1 Voice recognition and learning in L2 listeners

Figure 5.1 illustrates the voice recognition accuracy of the participants in the Test phases of the four training days. As shown by Figure 5.1, listeners demonstrated improvement. This observation was corroborated in the statistical analysis: Day was a significant predictor of accuracy (β=0.223, SE=0.038, p<.001). The inclusion of the factor Day to the model significantly improved the model fit ($\chi^2(1)$=24.72, p <.001). Participants thus managed to learn the target voices and to improve their recognition scores in a four-day training period.

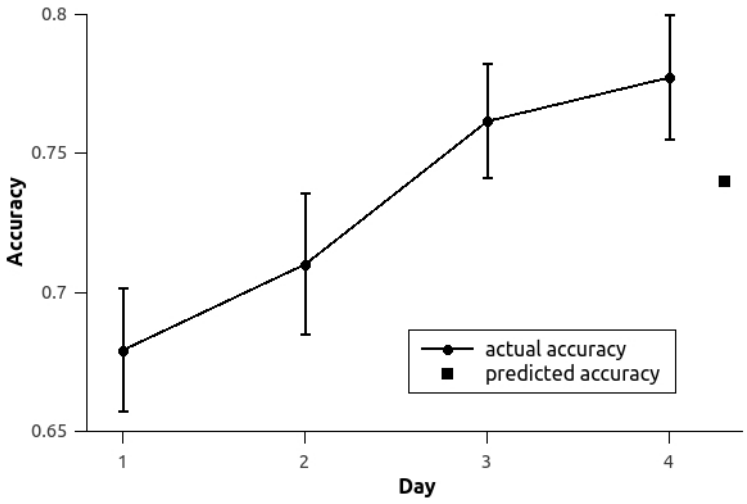

Figure 5.1: Voice learning performance across four consecutive days of training, averaged over all speakers and listeners (solid line with bullets), with error bars, and predicted accuracy on the basis of a multinomial regression analysis (black square; see explanation in the Section on speaker-related factors).

5.3.2 Speaker-related factors in voice recognition

In order to investigate to what extent acoustic voice characteristics play a role in voice learning and recognition, a measure of predictability of each voice based on its acoustics only was computed. According to earlier studies (see Baumann & Belin, 2010, for an overview), fundamental frequency (F0) is an important parameter for making judgments about the similarity between voices. To that end, the average F0, as well as the minimum and maximum F0, and the average word length per speaker (see Table 5.2) were submitted to a multinomial logistic regression in SPSS, with Speaker as the dependent variable. Based on the classification table output of this analysis, a new variable Predicted Accuracy was computed.

Predicted Accuracy is the percentage of times the voice was correctly predicted based on the F0 measures and the average word length. Since each participant was only trained on one specific list (the com-

bination of four speakers selected from the set of 12 speakers), the predicted accuracy was calculated for each speaker in each list of four speakers separately. Predicted Accuracy averaged across all lists is shown in Figure 5.1 (black square). As shown in Table 5.3 performance of the human listeners significantly correlated with the Predicted Accuracy (p <.01), with the highest correlation obtained for the third day of the training.

Table 5.3: Strength of the correlation between average participants' accuracy of voice recognition on each training day and predicted accuracy.

	Participants' accuracy			
	Day 1	Day 2	Day 3	Day 4
Predicted accuracy	0.45	0.58	0.61	0.5

Average performance of the listeners on Day 1 (M=67.83, SD=17.92) of the training was lower than the accuracy predicted on the basis of the acoustic parameters of the voice (M=74.38, SD=14.22), although this difference was only marginally significant (t(81.78)=-1.90, p = 0.06). On the last day of training, this difference reversed, but was again not significant (M=77.91, SD=12.37; t(84.37)=1.24, p=0.22). These results seem to suggest that participants succeeded using F0 and average word length to recognize the different voices from Day 1 onwards.

Predicted Accuracy for each speaker in a particular list was added as a factor to the overall analysis. The results showed that speakers that were recognized more accurately based on their acoustic characteristics by the computer were also recognized better by the human listeners (β=0.338, SE=0.047, p <.001) but at the same time Predicted Accuracy did not modify learning (the interaction Day x Predicted Accuracy was not significant). Moreover, the random factor List was excluded from the model, since it was no longer significantly improving model fit. The new model, including Predicted Accuracy and Day as fixed factors, Subject and Item as random factors (without List), and by Subject random slope for Day was a significant improvement over the earlier model (χ^2(0)=50.252, p <.001).

5.3.3 Stimulus-related factors in voice recognition

To investigate the role of the amount of indexical information in the word and lexical frequency on voice recognition, an indexical information measure and the frequency value on the Zipf scale (log10(frequency per million words) + 3) from the British SUBTLEX-UK word frequency database (Van Heuven et al., 2014) were added as fixed factors to the model of the previous subsection, as well as their interactions with the factor Day. We investigated three possible instantiations of the indexical information measure: number of syllables in the word (= number of vowels), number of nasals and vowels in the word (see the Introduction; Amino & Arai, 2008; Eatock & Mason, 1994; Gallardo et al., 2015), and length of the word (= number of phonemes). To avoid multicollinearity, these factors were not included in the same model. Rather, the models including only one of the three factors were compared with one another to establish which of these factors accounted for more variation in the data. All three measures turned out to be significant predictors of general accuracy in voice recognition, and none of them interacted with the factor Day. The model including the number of nasals and vowels had the lowest AIC (1637.4 against AIC=1638.6 for the model including number of syllables as a predictor, and AIC=1640.2 for the model including number of phonemes as a predictor). Therefore, the number of nasals and vowels in the word was included as a predictor in the subsequent analysis.

The results of the statistical analysis showed no significant effect for lexical frequency in voice recognition, not as a main effect nor as an interaction effect with the factor Day. Inclusion of the factor Frequency in the model (β=-0.021, SE=0.026, p=0.425) did not significantly improve model fit ($\chi^2(1)$=0.611, p=0.435). The best-fitting model, at this stage, therefore included Day, Number of nasals and vowels (stimulus-related) and Predicted Accuracy (speaker-related) as fixed factors, Subject and Speaker as random factors, and a by-Subject random slope for Day. The estimates of the fixed effects for this model are presented in Table 5.4. The stimulus-related random factor Word no longer contributed significantly to the model fit ($\chi^2(1)$=1.519, p=0.218) and was therefore removed.

Table 5.4: Estimates of the best-fitting model to predict voice recognition accuracy including all significant speaker- and stimulus-related factors.

Factor	β	SE	$p<$
Day	0.223	0.039	.001
Predicted accuracy	0.337	0.047	.001
Number of vowels and nasals	0.093	0.024	.001

5.3.4 Listener-related factors in voice recognition

Table 5.5 provides an overview of the scores for the two linguistic and one cognitive test, averaged over all listeners. The average score for LexTALE falls within the range of 60%-80% which corresponds to a B2 or upper-intermediate level of proficiency according to the Common European Framework of Reference (Lemhöfer & Broersma, 2012). The maximum possible score for Llama-D test is 75, and the average score of the participants in the present study corresponds to an "average score" for this test as specified by Meara (2005). Following Neger et al. (2014), we measured percentage of correct trials in the backward Digit Span task rather than, e.g., the highest number of digit strings correctly reproduced, since some participants made errors on both trials with four digits, but reproduced trials with five or six digits correctly. Percentage of correctly reproduced trials was then a fairer measure of their performance.

Table 5.5: Participants' performance on the language and cognitive tests. Standard deviations are provided between brackets.

LexTALE	Llama	Digit Span
72.5 (15.7)	32.7(16.9)	68.1 (19.5)

To study the role of individual differences in linguistic and cognitive skills on voice learning, z-transformed scores for Llama, LexTALE, backward Digit Span and their interactions with the factor Day were included in the best-fitting model from the previous subsection. Since the data for the LexTALE test for one participant was missing, 39, rather

than 40, participants were included in this analysis. The initial model for the analysis of the role of listener-related factors in voice recognition is presented in Table 5.6.

Table 5.6: Estimates of the initial model of voice recognition performance including the significant speaker- and stimulus-related factors and all individuals' linguistic and cognitive measures.

Factor	β	SE	p
Day	0.224	0.038	$<$.001
Predicted accuracy	0.336	0.048	$<$.001
Number of vowels and nasals	0.100	0.025	$<$.001
Llama	-0.050	0.122	0.686
Digit Span	0.199	0.121	0.100
LexTALE	0.094	0.117	0.421
Day x Llama	0.023	0.039	0.547
Day x Digit Span	-0.072	0.038	0.059
Day x LexTALE	0.029	0.038	0.436

As can be seen in Table 5.6, neither the LexTALE nor the Llama score played a role in accurate voice recognition and learning from Day 1 to Day 4 of the training. Digit Span and its interaction with Day were however marginally significant. After step-wise removal of all non-significant factors and interactions, this interaction remained marginally significant (β=-.004, SE=.002, p=.071). Although the removal of the interaction Day x Digit Span from the model increased the AIC and log likelihood of the model, the difference between the models with and without the interaction Day x Digit Span did not reach significance ($\chi^2(1)$=3.131, p=.077). Removal of the factor Digit Span also did not significantly decrease model fit ($\chi^2(1)$=.588, p=.443). These results seem to suggest that none of the participants' characteristics, such as lexical knowledge in the non-native language, phonetic aptitude, or Central Executive influenced voice recognition and learning.

The presence of the marginally significant interaction between Digit Span scores and Day might however indicate a different pattern of learning for the participants with high and low Digit Span scores. This hypothesis was further investigated by dividing the participants into two groups

according to their Digit Span score (range 33.3-100%; with 66.67% and higher belonging to the "high scores": 22 people). Figure 5.2 shows the voice recognition accuracy over the period of four training days for the participant group with the high backward Digit Span score and that of the participant group with the low backward Digit Span score.

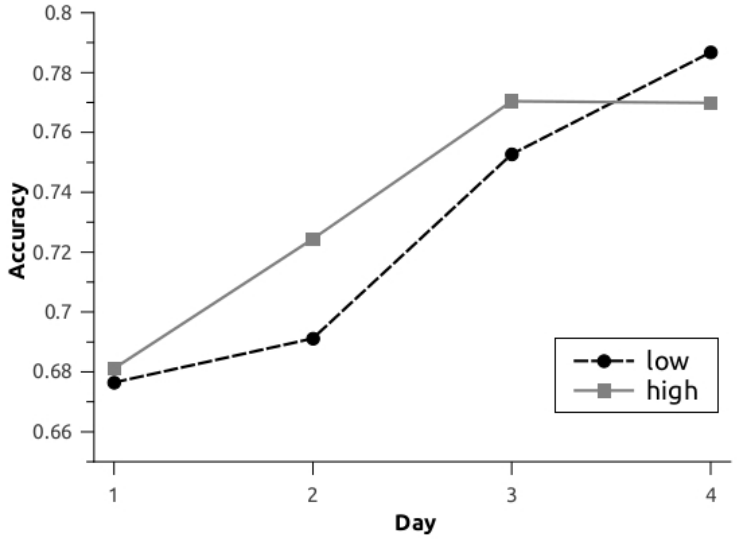

Figure 5.2: The rate of improvement in voice recognition accuracy for the groups of participants with a higher and lower Digit Span score. Solid line with squares represents performance of the participants with high backward Digit Span score. Dashed line with filled circles represents performance of the participants with low backward Digit Span score.

Figure 5.2 seems to suggest that people with lower working memory capacity learn more slowly than people with better working memory capacity (note that the accuracy score on the test on Day 1 was calculated after the familiarization and test phases thus after the first day of learning), but at the same time have more room for improvement. On the first days of learning, participants with a higher Digit Span score demonstrated a somewhat higher voice recognition accuracy than listeners with a lower Digit Span score, while on the last day this pattern reversed. Listeners with a higher Digit Span score seemed to have im-

proved their recognition more after the second day of the training and did not show any improvement from the third to the fourth day of the training, while listeners with a lower Digit Span score improved the most on the third day of the training. It appears, therefore, that working memory capacity influences voice learning speed. Nevertheless, since removal of the interaction did not significantly decrease the model fit, we have to be careful in interpreting these results.

5.4 Discussion

The present study investigates the combined and separate contribution of speaker-, listener-, and stimulus-related factors to voice learning and voice recognition in L2 listening. The results suggest that voice characteristics of the speaker, expressed in average, minimum and maximum F0, and the average word length in ms, as well as stimulus characteristics, i.e., the number of sounds in the word carrying indexical information contribute to L2 voice recognition performance. Interestingly, neither lexical frequency of an item nor lexical knowledge of the listeners (LexTALE score) played a role in voice learning and recognition performance. Moreover, no effect of phonological aptitude, expressed in the Llama test score, was observed. The results, however, seem to hint at a role of working memory capacity on the speed of learning to recognize voices.

Learning to recognize previously unfamiliar voices goes fast and seemingly effortlessly even for L2 listeners who arguably have less well developed lexical and phonological knowledge of the non-native language compared to native speakers of that language. The average voice recognition performance of the participants reached 68% correct already on the first day of the training, and significantly increased to 78% on the last day of the training. Factor Day (training) was a significant predictor of voice recognition accuracy in all the conducted analyses. The contribution of speaker-, stimulus-, and listener-related factors to the listeners' voice recognition and improvement in accuracy is discussed in detail below.

5.4.1 Speaker-related factors in voice recognition

The L2 listeners in the present study were shown to use the acoustics of the speakers' voices in voice recognition. There was a high correlation between the accuracy predicted by the multinomial logistic regression

analysis for the speaker for each list and the accuracy of the listeners. Moreover, voices that were better recognized in the multinomial logistic regression were also better recognized by the listeners. When comparing the predicted scores obtained with the classification table from the multinomial logistic regression analysis and the scores obtained by the human listeners (see Figure 5.1), we see that the accuracy scores obtained by the participants on the first day of training were lower than those of the regression analysis, but higher on the last day of the training. These results seem to point at an increase in listeners' sensitivity to speaker-specific acoustic characteristics due to the training. Voice learning thus seems to entail associating acoustic properties to particular voices, and doing so leads to higher recognition scores over time.

At the same time, listeners' better performance on Day 4 compared to the score predicted by the multinomial logistic regression and the lower correlation between the predicted accuracy and the accuracy scores of the listeners on Day 4 than on Day 3 seem to indicate that listeners use additional sources, not only low-level acoustic properties of the speech signal, to learn and identify voices. This observation corroborates findings of previous studies, showing that listeners are able to identify voices based only on their acoustic properties (Winters et al., 2008) and perform better if they are good at perceiving pitch differences (Xie & Myers, 2015) when the language is unfamiliar. However, when the language is familiar as in the case with the participants in the present study various additional, language-specific cues are used in voice recognition, while the ability to perceive differences in pitch no longer plays a role in voice recognition in a familiar language (Xie & Myers, 2015).

Speakers differed in how easy they were to recognize (see also Levi, 2014, who found a significant effect of speaker in listeners' recognition accuracy). Between-speaker differences can for a substantial part be explained by the acoustic characteristics of their voices. Speaker 5 was recognized best by the listeners in the present study (the recognition accuracy reached 94%, already after the first training day). This speaker had the lowest F0 in the set of speakers (see Table 5.2). Speaker 6, on the other hand, had the lowest accuracy score of all speakers on the last day (the recognition accuracy never got above 66%), but also had prototypical F0 measures. These findings are in line with the norm-based or prototype-based view on voice identity (Papcun et al., 1989), which states that voices that are more distant from the prototype are easier to remember for listeners. This theory was further developed by Belin

and colleagues (Baumann & Belin, 2010; Latinus & Belin, 2011; Yovel & Belin, 2013), who used a multi-dimensional voice space (with F0 being the primary dimension for voice-similarity judgments), in which all other voices are encoded relative to the prototypical voice. Not only voices more distant from the prototype are recognized and memorized better (Andics, 2013), they also induce greater neuronal activity in voice-sensitive cortex than more prototypical voices (Latinus, McAleer, Bestelmeyer, & Belin, 2013).

5.4.2 Stimulus-related factors in voice recognition

Vowels and nasals are the sounds that carry the largest amount of indexical information (Eatock & Mason, 1994). Consequently, we predicted that listeners would be more accurate in recognizing a speaker's voice when the speaker produced a word containing a higher number of vowels and nasals. The accuracy of voice recognition was indeed found to be higher for these words. This effect of the phonetic content of the utterance (i.e., number of vowels and nasals) on listeners' voice recognition performance shows that phonological and speaker-specific information interact in speech perception. Previous studies (see Amino & Arai, 2008 for an overview), demonstrated that speaker-specific physiological properties are reflected to a different extent in different speech sounds, which influences voice identification in native listening. Our results show that the sounds in the speech signal may enhance L2 voice recognition accuracy as well. Moreover, as suggested by Winters and colleagues (Winters et al., 2008), when listening in a non-native language, vowel categories specific to the native language of the listeners might be used to a larger extent in voice recognition. Although not directly tested in the present study, this could be an interesting question for further research.

Previous results on the role of lexical frequency in voice recognition are unclear. We hypothesized that lexical frequency might play a role in voice recognition by L2 listeners since this group of listeners is assumed to have weak(er) lexical representations, of low-frequency words in particular, and therefore might be less able to exploit acoustic-phonetic details to successfully recognize voices, similar to the children in the studies by Levi and Schwartz (2009) and Levi (2015). Lexical frequency, however, turned out not to play any role in L2 voice recognition and learning in the current experiment. This outcome corresponds however to the outcomes for native adults (Levi & Schwartz, 2013; Winters et al., 2008). These

results seem to suggest that lexical information is indeed not necessary for successful voice recognition (in line with Winters et al., 2008; Levi & Schwartz, 2013). On the other hand, it is also possible, as suggested by Perrachione and Wong (2007), that in a more familiar language (or in more familiar words), listeners are better able to exploit phonological and acoustic cues to differentiate between voices. If we assume that L2 adults, similar to native children, are better able to exploit phonological information in high-frequent (highly familiar) than in low-frequent (low familiar) words (White et al., 2013) then the absence of the effect of lexical frequency in the present study could be connected to the (relatively) high lexical knowledge of the L2 listeners and the small range but relatively high word frequencies for the materials used, so that even the "low-frequent" words were familiar to the listeners.

5.4.3 Listener-related factors in voice recognition

Different from previous voice-learning studies with L2 listeners, the present study included a measure of language proficiency (LexTALE) to investigate the role of lexical knowledge during voice learning and recognition in L2. The LexTALE score was not shown to modulate voice recognition accuracy of the listeners nor their learning over time. Hence, in the present study, lexical knowledge did not seem to play a role in voice recognition. At the same time, all listeners scored relatively high on the LexTALE test and the word frequency of the stimulus items was fairly high (see also the previous section), which could have allowed them to successfully exploit acoustic-phonetic cues available in the stimuli for learning the voices.

Given the availability of phonological information in the signal and earlier findings of listeners being able to exploit it for voice recognition (Zarate et al., 2015), as well as the hypothesis introduced by Levi (2014) that learning L2 sound categories and voice learning are connected, it is perhaps surprising that no effect of phonological aptitude was observed in the present study. Previous studies that found a facilitating effect of phonological memory on voice recognition, however, did not use the Llama-D task as used in this study, but either employed the Comprehensive Test of Phonological processing, including Memory for Digits and Non-Word repetition task (Perrachione et al., 2011) or the auditory verbal forward Digit Span task (Bregman & Creel, 2014; Levi, 2014) as a measure of phonological memory. The difference between the Llama-

D test and these measures is that while the verbal forward Digit Span and non-word repetition tasks tap into short-term phonological memory, Llama-D taps into the recognition memory for phonological sequences and long-term knowledge of phonological regularities which results from that. Speciale, Ellis, and Bywater (2004) demonstrated that it is the combination of memory for phonological sequences (measured in our study with Llama-D) and short-term phonological memory capacity (measured with forward Digit Span) that predicts both productive and perceptive L2 knowledge rather than short-term phonological memory alone, which implies that these two cognitive skills are not the same. Taken together, where individuals' short-term phonological memory seems important for learning voices, recognition memory for phonological sequences is not.

We observed a marginal, though interesting, effect of working memory capacity (more specifically, Central Executive: the ability of the listeners to simultaneously store and process information: Levi, 2014) on voice learning, which suggests a connection between working memory capacity and voice learning speed. On the first days of learning, participants with a lower working memory span demonstrated a somewhat lower learning accuracy than listeners with a higher working memory span, while on the last day of the training, listeners with a lower backward Digit Span score outperformed those with a higher backward Digit Span score. Interestingly, similar to our finding, in the voice learning study by Levi (2014) with school children, a significant negative effect of the score of backward Digit Span on the voice recognition performance was observed on the last day of training (in the general analysis and analysis of the performance on the first day of the training backward Digit Span did not reach significance). Levi (2014) explained these results by suggesting that listeners with higher backward Digit Span scores used a different strategy in voice recognition and learning. The data from the current study however seem to suggest a difference in speed of learning rather than the use of a different strategy between listeners with larger and smaller working memory capacity. Since this result was only marginally significant, more research is needed to determine the precise role (if any) of working memory in voice recognition of adult listeners.

The observation that both lexical knowledge and lexical frequency of items play no role in L2 voice recognition suggests that lexical information is not required for successful voice recognition and learning. This finding corresponds to the idea put forward by Mullennix, Pisoni, and Martin (1989) that information about speakers' voices is related to early

perceptual processes, namely, the extraction of the acoustic-phonetic information from the speech. These processes occur at the prelexical, rather than the lexical level of speech processing (Andics, 2006, 2013). If the lexical level had been involved in voice recognition, the role of lexical frequency should have been observable, since the effect of lexical frequency occurs at the earliest moment of lexical access (Dahan, Magnuson, & Tanenhaus, 2001). This, therefore, implies that (L2) voice recognition operates at the prelexical level of processing. This finding is in line with other studies, showing that while access to phonological information facilitates voice recognition, lexical and semantic access are not necessary for successful recognition of voices (Perrachione et al., 2011; Zarate et al., 2015).

The present study is the first one to investigate the combined role of speaker-related, listener-related, and stimulus-related factors in voice recognition and learning by L2 listeners. We have shown that L2 listeners are able to learn to recognize speakers' voices while they are speaking in a language that is non-native though familiar to the listeners, and that speaker-related, listener-related, and stimulus-related factors have an accumulative effect on voice recognition. Voice recognition is better for speakers whose acoustic characteristics are more deviant from the prototype (e.g., the lowest F0 in the set), in words which contain more indexical information (words with a larger number of vowels and nasals), while working memory capacity seems to influence the speed with which listeners learn to associate acoustic properties with specific speakers.

CHAPTER 6

Talker familiarity benefit in non-native speech processing and word recognition?

This Chapter is based on
Drozdova, P., van Hout, R., & Scharenborg, O. (2017). Talker familiarity benefit in non-native speech processing and word recognition? Manuscript submitted for publication.

Abstract

Native listeners benefit from talker familiarity in speech perception and word recognition, especially in adverse listening conditions. The present study addresses a talker familiarity benefit in non-native listening. Dutch listeners were trained to identify four English talkers over four days. The talker familiarity benefit in speech processing was investigated using a recognition memory task with old and new words by familiar and new talkers. The talker familiarity benefit in word recognition was studied by comparing performance on the first and the last day between the groups conducting the task in a familiar or in an unfamiliar voice. Non-native listeners demonstrated a talker familiarity benefit in speech processing, which was modulated by listening conditions, degree of familiarity with the voice, and non-native proficiency. No additional benefit of familiarity with the voice was found in word recognition. These results suggest that, similar to native listening, both abstract and talker-specific information influence non-native speech perception.

6.1 Introduction

The speech signal is extremely variable and contains not only linguistic but also so-called indexical information (Abercombie, 1967), i.e., for instance, information about a talker's age, gender, emotional state, dialect and accent. Previous studies have shown that talker-specific information is not ignored by listeners during speech perception. Words are better recognized when they are spoken by one talker than when they are spoken by multiple talkers (Mullennix et al., 1989; Ryalls & Pisoni, 1997). Moreover, words repeated by the same talker and even words combined from phonemes repeated by the same talker are recognized more quickly and more accurately than words repeated by a different talker, indicating that surface details of words, such as talker-specific information, are retained in some form in the memory of the listeners and facilitate speech processing at multiple levels (Jesse, McQueen, & Page, 2007; Luce & Lyons, 1998; Palmeri et al., 1993). Palmeri et al. (1993) argue that the benefit of a same-voice repetition is caused by a match in surface details between the information stored in listeners' memory and the repeated word. If there is a full match in both linguistic and indexical information, processing is facilitated.

Goh (2005), using a recognition memory task, demonstrated that listeners can take advantage not only of the full match between the surface characteristics of the first and second presentation of a stimulus (also known as the same talker benefit) but also of the familiarity with the voice of the speaker (i.e., the familiar talker benefit). Listeners in that study were more accurate at correctly indicating whether a word had already been presented earlier in the experiment not only when there was a complete match between the first and the second presentation of a word but also when the second presentation of the word was produced by a different talker whom the listener had already heard in the first part of the experiment, but who had not produced the target word, compared to when the second presentation of the word was produced by a completely new talker.

Talker familiarity has also been shown to play a role in word recognition. Nygaard and Pisoni (1998) trained listeners over the course of ten days to recognize the voices of previously unfamiliar talkers. The participants who were able to learn the voices were better at recognizing words in noise spoken by the familiar talkers (i.e., those they were trained to identify) than the same words in noise spoken by new talkers. This ef-

fect of familiarity with the voice of the talker was more pronounced for the most difficult noise levels, and was observed for both isolated words and sentences, not only for the words present during voice training but also for novel words (see also Nygaard et al., 1994). Similar findings were obtained with different types of voice training (Yonan & Sommers, 2000) and not only for young adults but also for children (Levi, 2014) and older adults (Yonan & Sommers, 2000). These studies provide additional evidence that listeners' memory retains not only linguistic, but also talker-specific information, and that there is a link between perceptual learning of talker identity and speech processing (Nygaard & Pisoni, 1998).

The presence or absence of the same talker benefit in speech processing was shown to be dependent on the nature of the task. The same talker effects were found using a recognition memory task (Goh, 2005; Goldinger, 1996; Luce & Lyons, 1998; Palmeri et al., 1993) and a stem-completion task, which requires listeners to complete the stem of the word to form any word that comes to mind (Schacter & Church, 1992). At the same time, Schacter and Church (1992) failed to observe a facilitatory effect of a same talker repetition using a long-term priming paradigm and cued-recall tasks which require listeners to complete the beginning of words which they heard earlier in an exposure phase, while Luce and Lyons (1998) did not observe any difference between same and different-voice presentations in a lexical decision task. The familiar talker benefit, on the other hand, was observed using a recognition memory task Goh (2005), as well as a word recognition task. The presence of a familiar talker benefit in word recognition was shown to be dependent on the noise level: with the highest benefit observed for the most difficult noise levels (Nygaard & Pisoni, 1998; Yonan & Sommers, 2000).

According to the time-course hypothesis (McLennan & Luce, 2005), a reason for the discrepancies between studies could be that the effects of indexical information emerge relatively late in processing. The benefit of matching voice information between the exposure phase and the test is argued to be due to processing speed. This hypothesis is corroborated by findings that when processing is slowed down through an increase in the difficulty of the task, indexical information influences speech processing in tasks where talker effects are typically not observed. For instance, while talker effects are typically not observed in lexical decision tasks (e.g., Luce & Lyons, 1998), when the task difficulty was increased through the use of word-like non-word stimuli McLennan & Luce (2005),

dysarthric (Mattys & Liss, 2008), foreign-accented speech (McLennan & González, 2012), or low-frequency words (Luce & Lyons, 1998), talker effects did emerge. Maibauer et al. (2014) suggested an extension to the time-course hypothesis. Their attention-based account hypothesized that when listeners pay extensive attention to the stimuli, indexical effects emerge even when processing is fast. Increase in attention can happen when hearing famous voices (Maibauer et al., 2014), when the focus is put explicitly on voices of the speakers (Theodore et al., 2015) or when taboo words are used (Tuft et al., 2016). Another way to slow down speech processing and recognition and increase listeners' attention is to present words in noise. Indeed, the same- and familiar talker advantages were observed in tasks in which words were embedded in noise (Goldinger, 1996; Nijveld et al., 2015; Nygaard & Pisoni, 1998; Nygaard et al., 1994; Yonan & Sommers, 2000). These studies suggest that degraded stimuli can increase talker-specific effects, arguably either through slowing down processing to the extent that indexical information can be accessed and used during speech processing or through making the listeners pay more attention to the acoustic-phonetic characteristics of the stimuli.

While native listeners have been repeatedly shown to benefit from the match in indexical characteristics between exposure and test, studies investigating the effect of indexical information on non-native speech processing and word recognition are scarce. Perceiving speech in a non-native language is more difficult than in a native language due to the mismatch in sound categories between the native and non-native language of the listeners, and consequently, the spurious activation of candidate words from both the native and non-native language during speech recognition (e.g., Broersma, 2012; Weber & Cutler, 2004). Moreover, listening in the presence of noise is more challenging for non-native than for native listeners (e.g., Mayo et al., 1997; Rogers et al., 2006; Scharenborg et al., 2017). On the one hand, it is possible that non-native listeners will have difficulty picking up indexical information from the signal due to their impaired perception. On the other hand, given that indexical information was shown to affect speech processing and recognition particularly when listening conditions are difficult, non-native listeners could benefit from storing information about the talker, similar to what has been found for native listeners.

Indeed, non-native listeners have been shown to be sensitive to indexical information in the speech signal and to recognize words better when the talker is held constant than when the talker changes from word to

word (Bradlow & Pisoni, 1999). Moreover, non-native listeners demonstrated a same talker benefit in a word-repetition task, being faster at repeating already presented words than new words but only when these words were produced by the same talker as during the exposure phase (Trofimovich, 2005). This facilitation was shown to be dependent on the amount of experience with the non-native language of the listeners (Trofimovich, 2008). Listeners with a longer length of residence, and, arguably, more extensive experience with the non-native language were more likely to demonstrate a same talker processing benefit. The author argued that more experienced listeners were more sensitive to phonetic detail in spoken words in the non-native language, and therefore, experienced more facilitation from the same voice than the listeners with less experience in the non-native language. Furthermore, Winters et al. (2013) showed that native English learners of German correctly recognized more German words as "old" when they were repeated in the same voice than when they were repeated in a different voice. These studies thus show that non-native listeners store indexical information in memory and have a same talker benefit in speech processing. The amount of this benefit is, however, likely to depend on the characteristics of the listeners.

While only one study to our knowledge directly investigated the native familiar talker benefit in speech processing (Goh, 2005), no previous studies addressed the non-native familiar talker benefit for speech processing and word recognition. One study, however, compared word-recognition performance of listeners trained to recognize voices in either their native or an unfamiliar language and tested in their native language with the voices on which they were trained and unfamiliar voices (Levi et al., 2011). Levi et al. (2011) showed that the familiar talker benefit depends on the language knowledge of the listeners. Listeners were trained to identify English-German bilingual speakers when they were speaking either in English (the native language of the listeners) or in German (which was an unfamiliar language to the listeners). When performing a subsequent word recognition task with the same (as learned in the either English or German training phase) and new speakers in English, only those listeners who were trained in English demonstrated a familiar talker benefit when recognizing words embedded in noise. Levi and colleagues (2011) concluded that listeners need to establish acoustic-phonetic links between talker information and what is being said during the training in order to benefit from talker familiarity in word recognition. Listeners with knowledge of the non-native language (non-native

listeners; in contrast to the listeners in the Levi et al. study who did not have any knowledge of German) are expected to be able to establish this connection. Moreover, when listening in the presence of background noise, non-native listeners have been shown to rely more on acoustic details than on lexical knowledge (Mattys et al., 2010). This reliance on acoustic details could allow them to pay attention to the acoustic characteristics of a speaker's voice, and as a result, non-native listeners might demonstrate a familiar talker benefit during word recognition in the presence of background noise.

The aim of the present study is to investigate the effect of a familiar voice on non-native speech processing and word recognition in both clean and noisy listening conditions. Focusing on non-native listening also allows us to investigate the possible effects of non-native language proficiency on talker familiarity effects. Specifically, we aim to answer the following research questions: 1) Do non-native listeners demonstrate a familiar talker benefit in both speech processing and word recognition? 2) Does this familiar talker benefit increase with increasing talker familiarity? 3) What is the effect of the presence of background noise on the familiar talker benefit? 4) What is the role of listener's proficiency in the non-native language on the emergence of the familiar talker benefit?

Non-native Dutch listeners of English were trained to recognize four previously unfamiliar British English speakers over the course of four days. Speech processing was studied by means of an explicit recognition memory task (an Old/New task) where words were presented in clean and in noise. The effect of talker familiarity on word recognition was studied in a word recognition task with various levels of noise. On the first experimental day, listeners' non-native proficiency was measured. Following the studies showing non-native listeners' sensitivity to indexical information (Bradlow & Pisoni, 1999; Trofimovich, 2005; Winters et al., 2013), we hypothesize that Dutch listeners will demonstrate a familiar talker benefit in both speech processing and word recognition. In line with native listeners, we expect an increase in the benefit as voices become more familiar (Maibauer et al., 2014) and predict a higher benefit for words in noise (Nijveld et al., 2015; Nygaard & Pisoni, 1998) and for listeners with a higher lexical proficiency (Trofimovich, 2008).

6.2 Method

6.2.1 Participants

Thirty five native Dutch participants (8 males, M_{age} = 22.4, SD_{age} = 2.34) took part in the experiment. Participants were recruited from the Radboud University Nijmegen subject pool and received either course credits or a monetary reward at the end of the four-day experiment. Prior to the experiment, all participants had to fill in a questionnaire containing questions about their hearing or possible learning disorders and difficulties they might experience when listening to speech in the presence of background noise. Only those participants indicating no history of hearing or learning disorders were included in the experiment. The average LexTALE score, the test used to measure the listeners' proficiency in English, was 72.1 (SD=17.2) which corresponds to an upper-intermediate level of second language proficiency (Lemhöfer & Broersma, 2012).

Two additional groups of participants took part in two pre-tests. Fourteen native Dutch participants (2 males, M_{age} = 21.71, SD_{age} = 2.02) took part in a word recognition study to determine the appropriate noise levels for the experiment, while 16 other participants (2 males, M_{age} = 21.94, SD_{age} = 2.84) took part in the pre-test to check the design of the experiment and difficulty levels of the tasks. Both pre-tests are described in more detail in subsequent subsections. None of the pre-test participants took part in the main experiment.

6.2.2 Overall design of the experiment

Table 6.1 shows the overview of the experiment, with the different tasks listed for each of the four days. The different tasks and experiments will be explained in detail in the following subsections. The overall design of the experiment was as follows. Day 1 started with a word recognition task in which the participants had to recognize 60 words at three different signal-to-noise ratios (SNR) and in the clean. To investigate whether talker familiarity has an effect on word recognition, one group of participants had to recognize words spoken by Talker A, who was one of the four talkers the participants had to learn over the course of the four days (familiar talker condition), while the other group had to recognize words spoken by Talker B, who was not included in the voice

learning part of the experiment (unfamiliar talker condition). A second word recognition experiment was conducted at the end of Day 4, and the improvement from Day 1 to Day 4 for the two listener groups was measured. Over the course of the four days, participants were trained to learn the voices of four talkers, with a unique combination of talkers for each participant. On each day, the voice learning task was followed by a recognition memory task (Old/New task), which investigated the influence of talker familiarity on speech processing. In the Old/New task, half of the words were "old" for the participants (were already presented in the voice learning task) and half of the words were "new". Crucially, half of the words were spoken by talkers that the participants were trained on and half of the words were spoken by talkers that were unknown to the participants. Moreover, half of the words in this task were presented in +5 SNR noise and the other half in clean listening conditions. To investigate the level of proficiency in the non-native language, the LexTALE task was carried out at the end of Day 1.

Table 6.1: Overview of the experimental tasks per day and the number of words, noise levels, and voices involved in each task. The column 'Duration' denotes the total duration of the experimental session for each day.

Day	Tasks	Words	Noise Level	Talkers	Duration
1	Word recognition	60	-5,0,5 SNR+clean	A/B	45 min
	Voice learning	24+128	clean	A, C, D, E	
	Old/New	32+32	+5 dB SNR+clean	A,C,D,E + 2 new talkers changed every day	
	LexTale				
2	Voice learning	128	clean	A, C, D, E	30 min
	Old/New	32+32	+5 dB SNR+clean	A,C,D,E + 2 new talkers	
3	Voice learning	128	clean	A, C, D, E	30 min
	Old/New	32+32	+5 dB SNR+clean	A,C,D,E + 2 new talkers	
4	Voice learning	128	clean	A, C, D, E	45 min
	Old/New	32+32	+5 dB SNR+clean	A,C,D,E + 2 new talkers	
	Word recognition	60	-5,0,5 SNR+clean	A/B	

6.2.3 Talkers

All stimuli were recorded by 13 male native British English speakers, who at the time of the experiment were living (working or studying) in or visiting the Netherlands. These are the same talkers as those who recorded the stimuli for the experiment described in Chapter 5. Moreover, an additional talker (Talker 13) was included, who recorded the stimuli which were used only in the word recognition task on the first and the last day of the experiment for one group of the participants.

Table 6.2 presents for each talker the average, standard deviation, and range (minimum and maximum) of the F0 as measured by Praat (Boersma & Weenink, 2009) in Hz, and the average word length in ms. Talker 1 was included in all tasks (the familiar talker in the word recognition task), Talkers 2-12 were used in the voice learning and Old/New tasks, while Talker 13 was only used in the unfamiliar talker condition in the word recognition task. In order to be able to compare the voice characteristics of the thirteen different talkers in the voice learning task, the average word length and F0 measures were calculated on the basis of the 128 words from the voice learning task. To compare the familiar (Talker 1; second mention in Table 6.2) and unfamiliar talker (Talker 13) in the word recognition task, the average word length and F0 measures were calculated on the basis of the 60 words from the word recognition task. Since it is important to compare Talker 1 both with the other talkers in the training set on the basis of 128 words and with Talker 13 on the basis of 60 other words, Talker 1 appears twice in Table 6.2.

As can be seen in Table 6.2, there was a variety of fundamental frequencies (ranging from 90 Hz to 179 Hz) and speaking rates (as indexed by the average word lengths which ranged from 431 ms to 624 ms). Since a unique combination of talkers was used for each participant in the voice learning and recognition memory tasks, talker familiarity effects could be investigated irrespective of how distant or similar the voices were. The role of the talker, the list in which the talker occurred during training, and predicted recognition accuracy for each talker in each list in voice-learning are investigated and described in detail in the previous Chapter.

The talkers were recorded individually in a sound-proof booth with a Sennheiser ME 64 microphone at a sampling frequency of 44100 Hz. Each word was pronounced at least twice by each speaker. Words which were mispronounced or produced too quietly were recorded again. The words

were then excised from the resulting audio files using a Matlab (The MathWorks Inc., 2013) script, and the segmentations were subsequently manually checked using Praat (Boersma & Weenink, 2009). All talkers were rewarded 5 Euro for half an hour of recording time.

Table 6.2: Characteristics of the talkers used in the experiment.

	F0				Word Length (ms)	
Talker	Mean	SD	Minimum	Maximum	Mean	SD
1	107	15	89	137	585	116
2	98	16	73	129	556	122
3	153	27	115	198	490	106
4	119	13	104	143	527	118
5	90	26	73	142	516	103
6	137	19	116	162	579	137
7	114	20	94	144	437	97
8	148	15	133	174	526	110
9	156	23	103	230	569	141
10	122	21	97	155	624	124
11	115	20	93	145	431	96
12	142	11	126	165	548	119
1	106	14	90	137	561	113
13	179	13	153	224	564	100

6.2.4 Materials, experimental set-up and procedure

All participants were tested individually in a quiet sound-attenuated booth. The stimuli were presented to them binaurally through headphones. The intensity level of all the stimuli was set at 70 dB SPL. The experiment was administered using Presentation software (Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com). In all tasks, participants were asked to react as quick as possible while trying to avoid making mistakes.

Voice learning task

In order to investigate the role of talker familiarity on speech processing and word recognition, it is crucial that participants familiarized themselves with the voices of the talkers used in the speech processing and word recognition tasks. To that end, a four-day voice learning task was implemented (see Chapter 5 for more details on the experimental set-up and an in-depth analysis of the results of the voice learning task). The voice learning task consisted of three parts: a familiarization, a voice training, and a test part, where the familiarization task was only present on the first day of the experiment.

The voice learning task included 76 monosyllabic and 76 bisyllabic content words (word frequencies ranged from 1.02 per million for *sob* to 589 per million for *end*). Because of a lexically-guided perceptual learning task not described in the current paper, none of the words used in the whole experiment contained /l/ or /r/ sounds. Twenty-four words (12 monosyllabic and 12 bisyllabic) were used in the familiarization task on Day 1. The remaining 128 words were used for the training and test tasks on Day 1-4. These words were semi-randomly split over the two tasks on each training day (so that each task contained the same number of bisyllabic and monosyllabic words and contained words of comparable frequency). The same words were used on each training day; however, following Nygaard and Pisoni (1998), the talker who produced the word on each day varied (i.e., if *day* was produced by Talker 1 on Day 1 it was, e.g., produced by Talker 2 on Day 2, etc.). Moreover, two different orders of presentation of the stimuli were used: the stimuli which were presented to half of the participants on the first day of the training were presented to the other half of the participants on the fourth day of the training, and the stimuli presented to half of the participants on the second day of the training were presented to the other half of the participants on the third training day, etc.

Each participant was trained to learn to recognize the voices of four talkers of the set of 12 talkers (Talker 1-12), where all participants had to learn the voice of Talker 1 (because of the word recognition task). To that end, 11 combinations of talkers (lists) were created (e.g., List 1: Talker 1, 5, 8, 9; List 2: Talker 1, 11, 7, 12 etc.). Each participant was randomly assigned to one of the lists.

In the *familiarization phase*, participants were instructed they would hear words spoken by four different talkers, and their task was to memo-

rize the voice and the name of the talker, which was shown on a computer screen. Participants were presented with five words from each talker, followed by a sequence of four words, each of which was again spoken by one of the four talkers. This procedure was repeated twice. Participants could press a button on a button box when they were ready to move to the next word.

In the *training phase*, participants saw the four names of the talkers on the screen. Upon hearing the stimulus, they had to press the button on the button box corresponding to the name of the talker they thought had spoken the word. Subsequently, the participants received feedback in the form of the word *correct* appearing on the screen in case of a correct response or the name of the correct talker in case of an incorrect response. The *test phase* was essentially the same as the *training phase* apart from the lack of feedback provided to the participants.

Old/New task

The Old/New task consisted of four conditions: Old talker/Old word, Old talker/New word, New talker/Old word, and New talker/New word. On each training day, the Old/New task included 64 words, 32 of which were already presented to the participants in the training or test phase of the voice learning task on that same day (16 from the training phase and 16 from the test phase), and 32 were new words. Note that of the old words spoken by the same talker as in the voice learning task, a different token (i.e., a different rendition) from the one used in the voice learning task, was chosen. The set of words was different for each training day. So, in total, 128 "old" words and 128 "new" words were used in the four Old/New tasks. The word frequencies in this task ranged from 1.02 per million for *sob* to 1778 per million for *back* (Van Heuven et al., 2014).

The second crucial manipulation was the speaker of the "old" and "new" words. Half of the words presented to the participants were spoken by the four talkers on which the listeners were trained, so that each "old" talker produced four "old" words (the same words but a different rendition compared to that in the voice learning part of the experiment on that day) and four "new" words. The other 32 words were spoken by two "new", unfamiliar talkers, so that each new talker produced eight "old" words and eight "new" words. A different set of words was used on each training day. Moreover, the two "new" talkers were different for each day. The new talkers were chosen from the remaining set of 8 talkers (12-

4 talkers on which the listener was trained). Finally, half of the words in each condition were presented in speech-shaped noise at an SNR of 5 dB. Addition of noise to the stimuli was done in the same way as described in the subsection *Word recognition task* below.

The participants were instructed that they would hear words, some of which they had already heard in either the training or the test phase of the voice learning task on that day. They were told that some words would be embedded in noise, but were asked not to pay attention to the noise or to the talker which produced the word, rather that they should only pay attention to the words themselves. The task of the participant was then to decide whether the word they heard was already presented to them in the voice learning part or whether the word was new. This task, therefore, required explicit recollection of previously heard items. Participants had to indicate their answer by pressing one of two buttons on the button box. To aid the listeners two options appeared on the screen: "old" corresponded to the left button and appeared on the left side of the screen and "new" corresponded to the right button and appeared on the right side of the screen.

Word recognition task

Thirty mono- and 30 bisyllabic content words were chosen from the SUBTLEX-UK database (Van Heuven et al., 2014). Word frequencies ranged from 0.2 per million for *skeptic* to 977 per million for *day*; the average frequencies for the monosyllabic and bisyllabic words were comparable.

Four listening conditions were used: one clean listening condition and three conditions with speech-shaped noise at three different SNRs. Each participant was presented with each listening condition and each word occurred only once in the experiment for each participant. To that end, the set of 60 words was divided into four blocks such that the number of monosyllabic and bisyllabic words and the word frequencies were similar in the four blocks. The listening conditions for each block were randomized across participants. Additionally, two different orders of presentation of the blocks (= two experimental lists) were used. Different renditions for each word by each talker were used on the first and fourth day of the experiment.

Following the procedure described in Scharenborg et al. (2017), noise at different SNRs was automatically added to the words using a PRAAT

script. Each word was preceded and followed by 200 ms of noise, and 20 ms of lead-in noise was added. Before adding noise the audio file was down sampled to 16000 Hz to match the sampling frequency of the noise file. Three different noise ratios were used: SNR=5 dB, 0 dB and -5 dB. These ratios were chosen on the basis of a pre-test, in which participants heard words from different talkers (Talker 1 and four other randomly chosen talkers from the set of 12 speakers) at different noise ratios (-10 dB, -5 dB, 0 dB, 5 dB) and had to type in the words they thought they had heard. Since -10 dB appeared to be too difficult for the listeners (overall accuracy below 20% correct), it was decided to use -5 dB as the lowest SNR (overall accuracy of 42%; 50% correct for Talker 1). Participants were randomly assigned to one of the experimental lists and to one of the two talker conditions, i.e., the familiar or unfamiliar talker condition. In the familiar talker condition, all words in the word recognition task were spoken by Talker 1 on which the participants were subsequently/previously trained in the voice learning part of the experiment. In the unfamiliar talker condition, all words in the word recognition task were spoken by Talker 13 who was not included in any of the other tasks in the experiment. The word recognition experiments on the first and last days had the same set-up but contained different renditions of each word. In the experiment, participants were instructed that they would hear words, some of which would be in noise, and they would have to type in the word they thought they had heard. Each block of 15 words was followed by a pause. To start the next block, participants had to press a key.

Language test (LexTALE)

Proficiency in the non-native language was assessed using a visual unspeeded lexical decision task for advanced learners of English (LexTALE: Lemhöfer & Broersma, 2012). Participants were presented with 60 words which were shown on a screen one-by-one. They had to indicate by button press whether the word on the screen was an existing word in English or not.

6.3 Results

Because of a technical error, the data from one participant on Day 3 for the Old/New task and from one participant on the Voice Learning task

were not recorded. Additionally, the LexTALE result of one of the participants was missing. The data of these three participants were excluded from the analyses where these measures were relevant.

Three sets of analyses were carried out, one for each task. In order to establish whether talker familiarity played a role in speech processing and word recognition, it is necessary to first establish whether participants indeed learned to recognize the four talkers, and specifically Talker 1. The first set of analyses investigated the responses in the voice learning task. Voice learning overall and of Talker 1 specifically should manifest itself as an increase in correct responses from the first to the last training day.

The second set of analyses focused on the responses and reaction times of the listeners in the recognition memory (Old/New) task to investigate the role of talker familiarity on speech processing. Crucially, the response patterns and reaction times were compared for items produced by the familiar and unfamiliar talkers. Following the studies focusing on the same talker and familiar talker advantage (e.g., Goh, 2005; Luce & Lyons, 1998; Palmeri et al., 1993), listeners were expected to be faster and more accurate recognizing items as old when they were produced by familiar than when they were produced by unfamiliar talkers. Furthermore, the role of background noise was investigated on the talker familiarity effect as well as whether the potential benefit of the familiar talker increases with increasing familiarity with the talker, and whether non-native language proficiency plays a role in the emergence of this benefit.

The third set of analyses investigated the effect of talker familiarity on word recognition by comparing the improvement in recognition performance of the listeners before and after voice training between the familiar and unfamiliar talker conditions. We expected listeners in the familiar talker condition to show more improvement than listeners in the unfamiliar talker condition.

6.3.1 Voice Learning

Participants' responses in the test phase of the voice learning task were analyzed using a repeated-measures analysis of variance (ANOVA; following Levi et al., 2011; Nygaard & Pisoni, 1998; Yonan & Sommers, 2000). Since each participant was exposed to the voices of four talkers, proportions of hits (correct responses) and false alarms (participant

thinks that the word was produced by the target talker while it was produced by another talker) were calculated. Since on each day participants identified 64 stimuli, the maximum number of correct responses for each talker was 16, while the maximum number of false alarms was 48. The proportions of hits and false alarms were used to calculate d-prime (d'), a common sensitivity index to measure accuracy performance of participants (Macmillan & Kaplan, 1985).

We were primarily interested in the improvement in the recognition accuracy over time (the within-subject factor Day). Due to the word recognition experiment (see subsection: *Talker familiarity effect in word recognition*: third set of analyses) being administered prior to the familiarization phase, listeners in the familiar talker condition had already been exposed to the voice of Talker 1 while the listeners in the unfamiliar talker condition were not. To account for this difference in exposure to the voice of Talker 1 and to investigate whether both listener groups were able to learn the voices of the talkers, Talker Condition (familiar vs. unfamiliar) was included in the analysis of the Voice Learning experiment as a between-subject factor.

Figure 6.1 illustrates the voice recognition performance in the test phase measured with d' averaged over all talkers, split out per training day and for the familiar and unfamiliar talker conditions separately.

The ANOVA analysis showed a significant difference between experimental days ($F(3,96) = 12.25$, $p < 0.001$). As can be seen in Figure 6.1, listeners improved their voice recognition performance from the first to the last experimental day. Neither the difference in performance between the talker conditions ($F(1,32) = 0.961$, $p = 0.334$) nor the Talker Condition by Day interaction were statistically significant ($F(3,96) = 0.57$, $p = 0.63$).

Figure 6.2 shows the Voice Learning results for Talker 1 only, since improvement in the recognition of this talker was important to account for the performance of the participants in the word recognition task (see subsection: *Talker familiarity effect in word recognition)*. The statistical analysis showed a significant improvement in the recognition of Talker 1 from Day 1 to Day 4 ($F(3,96) = 3.044$, $p = 0.033$), which again did not differ significantly for the listeners in the familiar and unfamiliar talker conditions in the word recognition experiment (Talker Condition was not significant, $F(1,32)=0.034$, $p=0.86$, nor the interaction between Talker Condition and Day, $F(3,96) = 1.767$, $p = 0.16$).

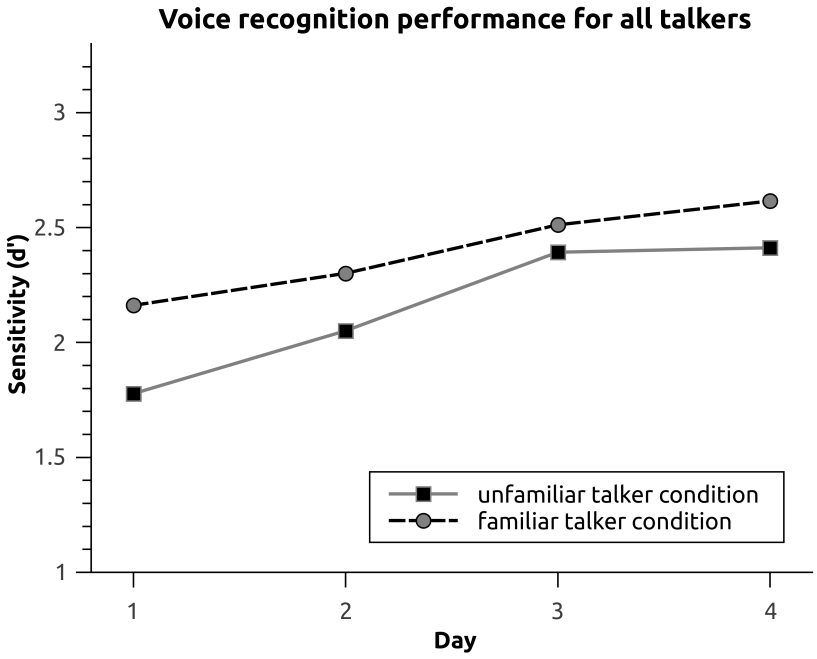

Figure 6.1: Voice learning performance or sensitivity (d') averaged across all words and talkers, split out per training day. Gray solid line with squares represents responses of the listeners in the unfamiliar talker condition. Black dashed line with bullets represents responses of the listeners in the familiar talker condition.

To summarize, the voice training was successful: participants improved the recognition of the four talkers they were exposed to as well as their recognition of Talker 1 irrespective of whether they had been exposed to the voice of Talker 1 prior to the Voice Learning experiment or not. A more detailed analysis of the voice learning results is beyond the scope of this article and can be found in Chapter 5.

6.3.2 Talker familiarity effect in speech processing: Old/ New task

To investigate whether talker familiarity facilitates speech processing in non-native listening, the sensitivity rates (expressed as d') for "Old"

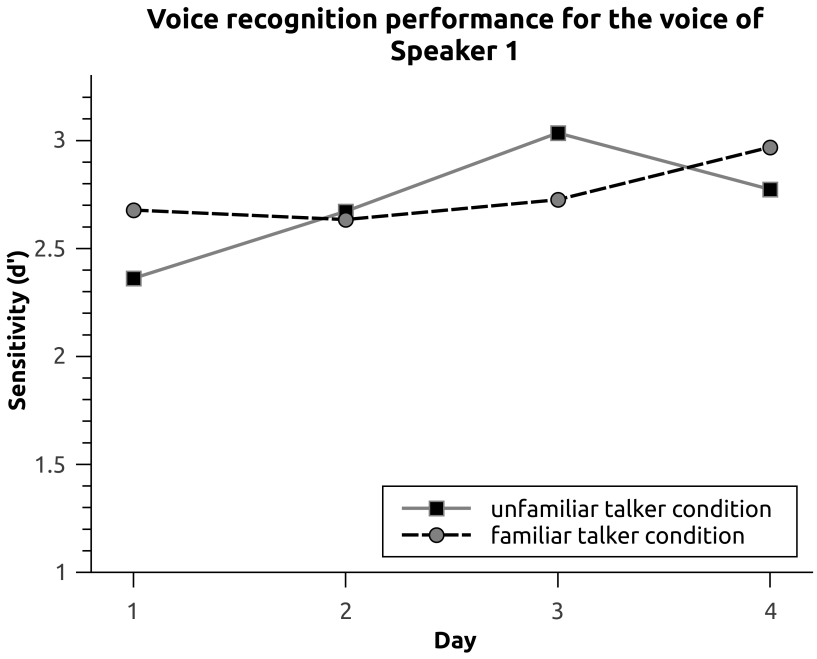

Figure 6.2: Voice learning performance or sensitivity (d') in recognizing the voice of Talker 1 for the two talker conditions, for each of the four training days.

words were computed for "Old" (familiar) and "New" (unfamiliar) talkers and the reaction times to hits were investigated (cf. Luce & Lyons, 1998; Palmeri et al., 1993; Goh, 2005). We also want to know whether this facilitation is modulated by the listening condition (clean vs. noise), the degree of sensitivity for recognizing the familiar talkers of the listeners (i.e., voice learning performance, see the previous subsection), and the non-native proficiency of the listeners.

All analyses for the Old/New task were conducted by means of linear mixed effects models (Jaeger, 2008). The analyses were performed in a step-wise manner starting from the most complex model including all the factors of interest (i.e., Day, Lexical Proficiency, Voice Learning Performance, Talker, and Noise) and the interactions between them.

Non-significant factors were removed from the model one by one, starting with non-significant interactions, and comparing each subsequent model with a previous one using the deviance score (- 2 * the log-likelihood ratio).

Accuracy

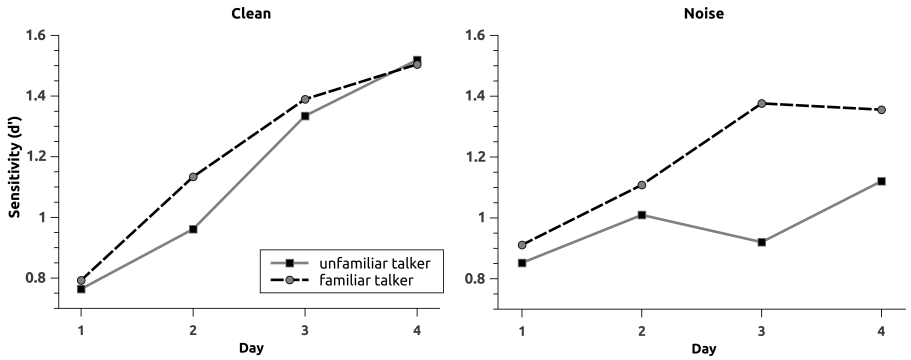

Figure 6.3: Sensitivity d' of the listeners in the Old/New task in recognizing old items, split out by Listening Condition and Talker (familiar vs. unfamiliar). The left panel shows the results for the clean listening condition; the right panel shows the results for the noisy listening condition.

The d' for the Old words was calculated for each participant per day and listening condition and included in the linear mixed effects model analysis as the dependent variable with Noise, Talker (familiar or unfamiliar), Day, Lexical Proficiency (non-native language proficiency measured with the LexTALE task) and the d' for Voice Learning Performance (Voice Learning) (described in the previous section) as fixed factors. Subject and List (combination of Talkers the participant was exposed to) were added as random factors. By-subject random slopes for Day, Noise and Talker were also included. Lexical Proficiency and Voice Learning were scaled and centered and Day was included as a categorical variable with Day 1 as reference value. The p values were obtained by treating the t statistics as z statistics (Barr, Levy, Scheepers, & Tily,

2013)[1]. Figure 6.3 illustrates the d' for recognizing old items produced by familiar (black dashed line with bullets) and unfamiliar (gray solid line with squares) talkers per day and per listening condition, with the results for the clean listening condition in the left panel and those for the noise condition in the right panel. The estimates from the best fitting model from this analysis are provided in Table 6.3. This final model only included Subject as a random factor, as other random slopes and intercepts were not shown to significantly improve the model.

Table 6.3: Estimates for the best fitting model for the d' measures of the old items in Old/New task.

Fixed effect	β	SE	t	p
Intercept	0.928	0.110	8.441	<0.001
Day 2	0.181	0.123	1.476	0.140
Day 3	0.496	0.125	3.968	<0.001
Day 4	0.610	0.126	4.832	<0.001
Talker	-0.136	0.058	-2.232	0.020
Noise	-0.005	0.127	-0.039	0.969
Lexical Proficiency	0.000	0.055	0.000	1.000
Voice Learning	0.184	0.080	2.289	0.022
Voice Learning x Noise	-0.279	0.108	-2.592	0.010
Day 2 x Noise	0.049	0.173	0.284	0.777
Day 3 x Noise	-0.189	0.175	-1.079	0.281
Day 4 x Noise	-0.193	0.176	-1.094	0.274
Lexical Proficiency x Noise	0.091	0.061	1.499	0.134
Day 2 x Voice Learning	-0.215	0.102	-2.103	0.035
Day 3 x Voice Learning	-0.170	0.114	-1.494	0.135
Day 4 x Voice Learning	-0.074	0.106	-0.697	0.486
Lexical Proficiency x Voice Learning	0.021	0.045	0.460	0.645
Noise x Lexical Proficiency x Voice Learning	-0.106	0.054	-1.943	0.052
Noise x Day 2 x Voice Learning	0.364	0.143	2.542	0.011
Noise x Day 3 x Voice Learning	0.286	0.159	1.795	0.073
Noise x Day 4 x Voice Learning	0.141	0.148	0.952	0.341

Importantly, the overall analysis for both listening conditions together revealed a general effect of talker familiarity (Table 6.3: Talker):

[1]Similar values were obtained when using a more conservative Kenward-Roger approximation

listeners were more accurate in recognizing or identifying old items produced by old, familiar talkers than by new, unfamiliar talkers, thus, demonstrating a familiar talker benefit. Another important finding is that the factor Noise is involved in many interactions (i.e., significant interaction between Noise and Voice Learning, and the three-way interactions between Noise, Voice Learning, Lexical Proficiency and between Noise, Voice Learning, Day). These interactions reveal systematic differences in performance of the listeners on words in clean and in noise. To investigate these differences between the clean and noisy listening conditions further, two new analyses were carried out, for the clean and noise listening conditions separately. Table 6.4 provides estimates for the best fitting model for the words presented in clean, and Table 6.5 provides estimates for the best fitting model for the words presented in noise.

Table 6.4: Estimates for the best fitting model for the d' measures of the old items in clean in the Old/New task.

Fixed effect	β	SE	t	p
Intercept	0.862	0.093	9.308	$<$0.001
Day 2	0.180	0.121	1.489	0.137
Day 3	0.493	0.123	4.006	$<$0.001
Day 4	0.608	0.124	4.907	$<$0.001
Voice Learning	0.185	0.078	2.383	0.017
Day 2 x Voice Learning	-0.221	0.101	-2.185	0.029
Day 3 x Voice Learning	-0.158	0.118	-1.410	0.158
Day 4 x Voice Learning	-0.066	0.105	-0.635	0.526

In the clean listening condition, listeners significantly improved their performance from Day 1 to Day 4 (see the increasing β values of Day 2 to Day 4 in Table 6.4). Moreover, listeners who were more accurate in identifying familiar voices were also better in identifying old words (factor Voice Learning). This effect is moderated by Day, implying that the Voice Learning effect was absent on Days 2 and 3, and present on Days 1 and 4. The analysis for the clean listening condition however did not reveal a significant difference in performance for familiar talker and unfamiliar talker (the Talker effect, see also the left panel in Figure 6.3 and Table 6.4). In other words, no direct familiar talker benefit was

observed. The Voice Learning effect indicates an indirect familiar talker effect as it means that an overall successful performance in recognizing familiar voices is related to higher scores in recognizing old items.

Table 6.5: Estimates for the best fitting model for the d' measures of the old items in noise in the Old/New task.

Fixed effect	β	SE	t	p
Intercept	0.988	0.101	9.795	$<$0.001
Day 2	0.177	0.120	1.476	0.140
Day 3	0.267	0.120	2.219	0.027
Day 4	0.356	0.120	2.966	0.003
Talker	-0.212	0.085	-2.498	0.012

The analysis of the noise listening condition showed a significant difference in performance for the words produced by the familiar and the unfamiliar talkers (see the right panel in Figure 6.3 and the Talker effect in Table 6.5), showing a familiar talker benefit. Again, participants improved their performance from Day 1 to Day 4. The increasing β values, however, are lower than in the clean condition, as already indicated by the Noise by Day interaction in Table 6.3. In both analyses, the random factor List and the by-Subject slopes for Day and Speaker did not significantly improve the model. The final models only included Subject as a random factor.

Reaction Times

Reaction times calculated from the word's offset that were more than two standard deviations from the mean or were below zero were removed, which resulted in the deletion of about 4.5% of the data. The outliers were calculated separately for the clean and noise listening conditions. Figure 6.4 shows the log transformed reaction times measured from the word's offset for the correct identification of the old words (hits) when the words were presented in the clean (left panel) and in noise (right panel). Again, responses of the listeners to the words pronounced by the old familiar talkers are shown with the black dashed line with bullets and responses to the words pronounced by the new unfamiliar talkers are shown with the gray solid line with squares.

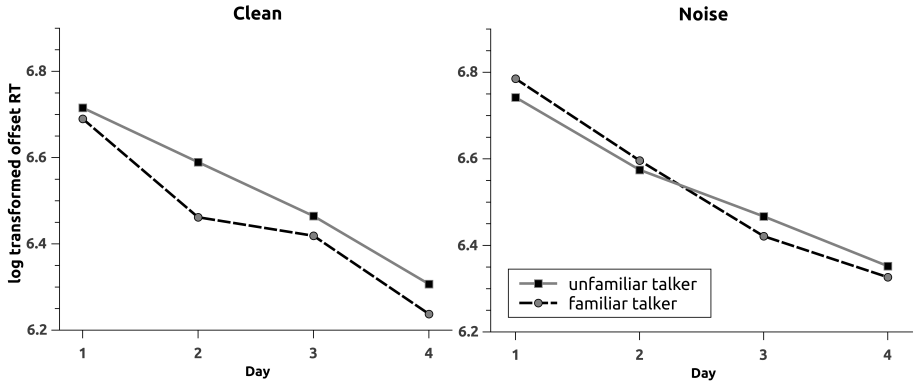

Figure 6.4: Response times for correct identification as old of the items spoken by the familiar (black dashed line with bullets) and new talkers (gray solid line with squares). The left panel shows the results for the clean listening condition; the right panel shows the results for the noisy listening condition.

Log transformed offset reaction times were used as the dependent variable in a linear mixed effects model analysis. The initial model included Talker (familiar or unfamiliar), Day (as a categorical variable with Day 1 as the reference), Noise, Voice Learning (scaled and centered) and Lexical Proficiency (scaled and centered) and all possible interactions between them as fixed factors. Subject, Item, Talker Number, and List were entered as random factors, with Subject random slopes for Noise, Talker, and Day. The estimates of the best-fitting model are presented in Table 6.6. The final best-fitting model only included Subject, Item, and Talker Number as random factors, and by Subject random slope for Day.

The initial analysis revealed a significant improvement in reaction times for both items in clean and in noise over time: participants became faster at giving correct responses to old items from the first to the last experimental day (see also Figure 6.4). Furthermore, participants were in general slower to react to the items in noise than to the items in clean. Importantly, there was a significant interaction between Talker and Lexical Proficiency, indicating that the difference in reaction times

Table 6.6: Estimates for the best fitting model for the reaction times of the hits for the old items in the Old/New task.

Fixed effect	β	SE	t	p
Intercept	6.602	0.074	88.85	<0.001
Day 2	-0.235	0.071	-3.33	0.001
Day 3	-0.359	0.069	-5.16	<0.001
Day 4	-0.567	0.057	-9.90	<0.001
Talker	-0.018	0.030	-0.27	0.787
Noise	0.087	0.026	3.23	0.001
Lexical Proficiency	0.003	0.050	0.06	0.951
Voice Learning	-0.023	0.029	-0.81	0.417
Day 2 x Talker	0.104	0.089	1.17	0.240
Day 3 x Talker	0.096	0.088	1.09	0.275
Day 4 x Talker	0.147	0.062	2.37	0.018
Lexical Proficiency x Talker	0.065	0.021	3.09	0.002
Voice Learning x Talker	0.022	0.025	0.90	0.367
Voice Learning x Noise	0.060	0.023	2.61	0.009
Noise x Talker	-0.040	0.040	-1.02	0.308
Noise x Talker x Voice Learning	-0.072	0.034	-2.13	0.033

to the words spoken by familiar and unfamiliar talkers was higher for listeners with a higher lexical proficiency than for listeners with a lower lexical proficiency. More proficient non-native listeners had a larger familiar talker benefit than less proficient non-native listeners. Moreover, there was a significant interaction between factors Talker and Day on Day 4, indicating that the difference between familiar and unfamiliar talkers increased on the last experimental day in comparison to the first experimental day. Talker also entered a significant three-way interaction with Noise and Voice Learning, indicating that the difference between words spoken by familiar and unfamiliar talkers was modulated by the voice learning performance of the listeners, depending on the listening condition. For the words in clean, larger voice Learning performance corresponded to larger differences in reaction times to the items by new and old speakers (with faster reaction times to the words by old speakers). For the words in noise, this tendency was reversed with lower voice learning performance corresponding to larger differences in reaction times to the items by old and new speakers. To investigate the differences between the

clean and noise listening conditions and the effect of Voice Learning on the familiar voice benefit in more detail, separate analyses were carried out for the clean and noise listening conditions. The best-fitting model for the reaction times for the words in clean is presented in Table 6.7, and the best-fitting model for the items in noise is presented in Table 6.8.

Table 6.7: Estimates for the best fitting model for the reaction times of the hits for the old items in the Old/New task in clean.

Fixed effect	β	SE	t	p
Intercept	6.555	0.071	92.14	<0.001
Day 2	-0.194	0.054	-3.575	<0.001
Day 3	-0.282	0.053	-5.294	<0.001
Day 4	-0.500	0.044	-11.486	<0.001
Talker	0.075	0.045	1.657	0.097
Lexical Proficiency	0.009	0.051	0.184	0.854
Lexical Proficiency x Talker	0.075	0.030	2.509	0.012

The best-fitting model for the reaction times for the items in clean and for the items in noise only included random intercepts for Subject, Item, and Talker Number in the random structure. The listeners significantly decreased their reaction times for the items in clean from Day 1 to Day 4 (left panel of Figure 6.4). Crucially, similar to the general analysis, a significant interaction between Talker and Lexical Proficiency was found, indicating that the difference between the items produced by familiar and unfamiliar talkers (faster reaction times to the items produced by familiar talkers) was only present for the listeners with a higher lexical proficiency.

The analysis of the items in noise showed that the listeners significantly decreased their reaction times for the words in noise as the experiment progressed (factor Day; right panel of Figure 6.4). No difference, however, was observed for the items produced by the familiar and unfamiliar talkers.

To summarize, the familiar talker benefit revealed itself in a higher accuracy for the old items (measured with d') when these items were produced by familiar talkers. This difference was only significant for noise,

Table 6.8: Estimates for the best fitting model for the reaction times of the hits for the old items in the Old/New task in noise.

Fixed effect	β	SE	t	p
Intercept	6.673	0.065	102.2	$<$0.001
Day 2	-0.192	0.052	-3.72	$<$0.001
Day 3	-0.362	0.051	-7.05	$<$0.001
Day 4	-0.499	0.042	-11.88	$<$0.001

although we had an indirect familiar talker effect by the interaction in clean between Voice Learning and Day. In the analysis of reaction times, the familiar talker benefit revealed itself in shorter reaction times for the items produced by familiar speakers in the clean listening condition when the listeners had higher lexical proficiency, while no effect of familiarity with the voice of the talker was observed for the reaction times on the items embedded in noise. Table 6.9 provides an overview of the results for the recognition memory (Old/New) task: "yes" indicates the presence of a talker familiarity benefit, while "no" indicates the absence of a (direct) talker familiarity benefit.

Table 6.9: Overview of the presence/absence of the talker familiarity effects in the Old/New task.

	All Items	Clean	Noise
Accuracy	Yes	No	Yes
Reaction Times	Yes	Yes	No

6.3.3 Talker familiarity effect in word recognition

In the final set of analyses, we investigated whether familiarity with a talker improved word recognition performance in noise and clean in non-native listeners. If so, this should manifest itself as a larger improvement in word recognition performance from Day 1 to Day 4 for the group of listeners from the familiar talker condition compared to the group of listeners in the unfamiliar talker condition. Moreover, we expect this

difference in improvement to be more prominent for the more difficult noise levels. Responses of the participants were coded as 1 if the answer was correct and 0 if the answer was incorrect. Obvious typing errors were corrected. The left panel of Figure 6.5 shows the word recognition accuracy for the familiar talker condition, while the right panel shows the word recognition accuracy of the unfamiliar talker condition.

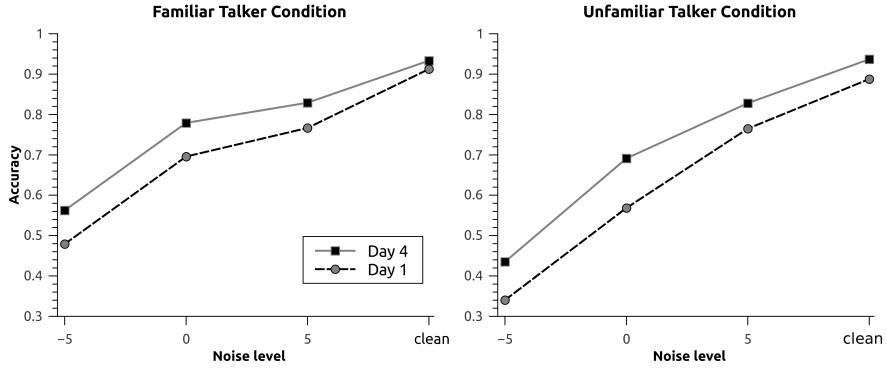

Figure 6.5: Word recognition accuracy of the two listener groups for the four noise conditions. The left panel shows the results for the familiar talker condition; the right panel shows the results for the unfamiliar talker condition. The black dashed line with bullets represents responses of the participants on the first training day. The gray solid line with squares represents responses of the participants on the last training day.

Figure 6.5 shows that the proportions of correctly recognized words increased as listening conditions became easier. Importantly, both participant groups seemed to perform better on Day 4 than on Day 1, so also the listeners who did not receive any training on the target voice showed an increase in word recognition accuracy (see right panel of Figure 6.5). At the same time, the plots show that on Day 1 the performance of the group of listeners who were familiarized with the talker was higher than that of the unfamiliar talker condition, which theoretically means that the participants in the familiar talker condition could improve less than those in the unfamiliar talker condition. We therefore calculated a measure of word recognition improvement that takes into account a lis-

tener's maximum possible improvement, which we refer to as "Relative Progress":

$$(a_2 - a_1)/(1 - a_1)$$

where a_1 is word recognition performance (proportion of correct responses) of the participant on the first training day and a_2 is performance of the participant on the last training day.

To investigate whether the listeners from the familiar talker condition demonstrated more progress than the listeners from the unfamiliar talker condition and whether this difference in improvement was modulated by the presence of noise, Relative Progress was used as the dependent variable in a linear mixed effects model analysis with SNR (-5, 0, 5, clean: with clean as a reference) and Talker Condition (familiar talker versus an unfamiliar talker) and their interaction as fixed factors. Additionally, lexical proficiency (i.e., the centered and scaled LexTALE score) was added as a fixed factor and in interaction with other factors, and Subject was added as a random factor. We expected to find a significant effect of Talker Condition and an interaction of Talker Condition with SNR. The estimates from the best fitting model are presented in Table 6.10.

Table 6.10: Estimates for the best fitting model of the word recognition task.

Fixed effect	β	*SE*	t	p
Intercept	0.292	0.085	3.437	$<$0.001
Lexical Proficiency	0.247	0.085	2.902	0.004
SNR -5	-0.167	0.115	-1.461	0.144
SNR 0	-0.092	0.115	-0.797	0.425
SNR 5	-0.162	0.115	-1.413	0.158
SNR -5 x Lexical Proficiency	-0.260	0.115	-2.260	0.024
SNR 0 x Lexical Proficiency	-0.291	0.115	-2.527	0.011
SNR 5 x Lexical Proficiency	-0.274	0.115	-2.375	0.018

Importantly, neither the interaction of Talker Condition with the other factors nor Talker Condition as a fixed effect significantly improved the model fit, showing that progress of the listeners did not depend on whether they received training on the target voice or not. This is

corroborated by the significant intercept which also indicates that the listeners' word recognition accuracy improves irrespective of whether they received voice recognition training on the target voice. In other words, no familiar talker advantage was observed. Participants with a higher LexTALE score demonstrated the most progress, however the role of lexical proficiency varied depending on the SNR (see the interactions between Lexical Proficiency and SNR): the difference in improvement between different noise levels was higher for the participants with lower lexical proficiency or in other words, listeners with a lower proficiency in the non-native language suffered more from deteriorating listening conditions.

To further probe this absence of a familiar talker advantage in the word recognition task, a second analysis was conducted, which included voice learning performance on the last training day as a potential predictor of Relative Progress (see also Levi et al., 2011). We included the d' score for Talker 1 on the last training day (see *Voice learning* subsection) of each listener as an indicator of Voice Learning in the final model from the previous analysis (see Table 6.10). If talker familiarity plays a role in non-native word recognition than the listeners from the familiar talker condition who are better at learning the target voice should demonstrate more relative progress in word recognition performance than listeners who are worse at learning the target voice, while for the listeners from the unfamiliar talker condition relative voice learning progress should not play a role. In other words, if familiarity with the talker plays a role in non-native word recognition, we expect to find Voice Learning influencing Relative Progress (more) for the familiar talker condition but not the unfamiliar speaker condition (i.e., a higher Relative Progress for the listeners with a higher d' score in the Familiar Talker condition). However, addition of the interaction Talker Condition and Voice Learning did not significantly improve model fit ($\chi^2(3)$=1.282, p=0.734). To summarize, no familiar talker benefit was observed in the word recognition task.

6.4 Discussion

The present study is the first study that investigated the effect of familiarity with a talker's voice on non-native speech processing and word recognition taking into account the presence of background noise. In addition, we investigated the impact of the degree of familiarity with the

voice of the talker and the non-native lexical proficiency of the listener in the non-native language. Following previous studies on the same talker benefit in non-native speech processing (Trofimovich, 2005, 2008; Winters et al., 2013) and on the familiar talker advantage in native speech processing (Goh, 2005) and word recognition (Nygaard et al., 1994; Nygaard & Pisoni, 1998), we hypothesized that non-native listeners enjoy the same familiar talker benefit in both non-native speech processing and word recognition, as rendered in the first research question. After successful learning to recognize four previously unfamiliar talkers over the course of four days, the advantage of using a familiar talker was only observed in the recognition memory task, which arguably taps into speech processing, but not in the word recognition task. Our results on the speech processing task are in line with those found for native listeners regarding the familiar talker advantage (Goh, 2005). Our results on the word recognition task, however, deviate from those found for native listening (e.g., Nygaard et al., 1994; Nygaard & Pisoni, 1998; Yonan & Sommers, 2000) but are in line with those found by Levi and colleagues (2011) who also did not observe a familiar talker effect in a word recognition task when listeners were trained to identify talkers in an unfamiliar, but phonotactically related, language.

Several explanations can be put forward for the different results on the familiar talker benefit for native and non-native listeners. While all previous studies (Nygaard et al., 1994; Nygaard & Pisoni, 1998; Yonan & Sommers, 2000) only included a word or sentence recognition task at the end of the experiment on the last day of the voice training, we used a pre- and post-test design in which performance of the listeners on the first experimental day (before the voice training) and the last experimental day (after the voice training) were compared. Since the same words, albeit different renditions of these words, were used in the pre- and post-tests, it is possible that a same talker advantage interfered with the familiar talker benefit in the present study. Goldinger (1996) showed that the same talker advantage in a word-in-noise identification task is present even after a one-week delay. The match in indexical and linguistic information between the first and last training day in our experiment could, therefore, in principle have led to an improvement in performance for both groups (i.e., the familiar talker condition and unfamiliar talker condition), thus showing a same talker advantage (albeit with different voices for the two groups). Newman and Evers (2007) suggested that even small amounts of exposure to a voice are sufficient to generate familiarity,

and that large amounts of exposure generate little additional benefit. Possibly, the exposure to the target voice on Day 1 (during the word recognition experiment) in the present study was enough for the speakers to familiarize themselves with the talker. The additional familiarization in the familiar talker condition did not lead to an additional benefit of familiarity with the talker in word recognition.

Levi et al. (2011) hypothesized that the mapping between acoustic information and linguistic information during voice training is crucial for a familiar talker benefit to emerge. The present study tested Dutch non-native listeners in English, who were hypothesized to be able to establish acoustic-phonetic links between talker information and what is being said during the voice training. Potentially, however, contrary to the native listeners tested by Levi et al. (2011), the non-native listeners in the present study might still have primarily relied on language-independent talker information during voice learning, such as pitch, and characteristic formant values, rather than lexical information, thus preventing the familiar talker advantage to arise. This explanation is, however, less likely since starting from the second day of the experiment, listeners found out that they had to pay attention to the words talkers were producing during voice training, since the exact same words were later used in the Old/New task.

A third explanation for the presence of the familiar talker benefit in the recognition memory task and its absence in the word recognition task is related to the differences between the two tasks and specifically the cognitive speech comprehension processes they tap into. While the word recognition task focuses on deeper lexical processing abstracting from voice specific characteristics, less deep processing and no abstraction of voice specific characteristics of spoken words is required in the Old/New task. Possibly, the observed differences might be connected to how indexical information is stored in the memory of the listeners. Cutler and colleagues (Cutler, 2010; Cutler et al., 2010) suggested that indexical information is not stored as part of the lexical representation of a word but rather in an episodic memory system separate from the mental lexicon. This episodic memory system is accessed in an Old/New task but not in a word-recognition task. This explanation corroborates with studies that found talker-related effects in the tasks relying on less deep processing (Goldinger, 1996; Luce & Lyons, 1998) but not in the tasks requiring lexical processing (e.g., no talker-related effects in a lexical decision task: Luce & Lyons, 1998, or a semantic priming task: Kittredge,

Davis, & Blumstein, 2006; Lee & Zhang, 2015.

The second research question addressed in this study focused on whether increasing talker familiarity has a positive effect on the talker familiarity benefit in speech processing and word recognition. Since we only observed a familiar talker effect in the Old/New task, we will primarily focus on the results of the Old/New task, i.e., the talker familiarity effect in speech processing. Here, increased familiarity with the voice led to faster reaction times: larger differences between words produced by familiar talkers and unfamiliar talkers were observed on the last experimental day than on the first experimental day which could indicate that the familiar talker benefit increases as the talker becomes more familiar. These results are in line with Maibauer et al. (2014) who suggested that listeners pay more attention to familiar than to unfamiliar talkers and that therefore indexical effects are easier to observe when the voices are familiar. Voice learning progress of the listeners was, however, not shown to influence the talker familiarity benefit, contrary to what was observed in previous studies with native listeners (e.g., Levi et al., 2011; Nygaard & Pisoni, 1998), which can theoretically be related to the generally high voice recognition accuracy of the listeners in the present study (68% correct already on the first training day).

The third research question in the present study addressed the role of background noise in the emergence of the familiar talker benefit. In line with numerous studies on non-native speech comprehension in the presence of background noise (e.g., Garcia Lecumberri et al., 2010; Scharenborg et al., 2017), the presence of background noise had a negative effect on speech comprehension. The results showed that participants were slower to react to items in noise than to items in clean in the Old/New task, while fewer words were recognized in worse listening conditions in the word recognition task. According to the time-course hypothesis (McLennan & Luce, 2005), a larger familiar talker benefit could be expected to occur for items in noise than in clean in the Old/New task. According to this hypothesis, the effect of talker-specific information emerges relatively late in processing. When processing is made relatively slow by increasing the difficulty of the task, e.g., the addition of noise, talker-specific effects are observed even in a lexical decision task, which is relatively fast and requires lexical access (Luce & Lyons, 1998; Mattys & Liss, 2008; McLennan & González, 2012; McLennan & Luce, 2005). Our findings do not fully support the time-course hypothesis. Although the presence of noise in the stimuli did influence the emergence of the

talker familiarity benefit, its effect differed depending on whether accuracy or reaction times were measured. When accuracy was analyzed the familiar talker benefit was observed for the words in noise but not for the words in clean which is in line with the time-course hypothesis. Note, however, that the combined analysis of the items in clean and in noise revealed a significant effect of familiarity with the speaker which was not modulated by listening conditions. This suggests that also in the clean listening condition listeners were more accurate at reacting to the items produced by familiar talkers than by unfamiliar talkers, but this difference was not large enough to reach significance. At the same time, in the reaction time analysis, the familiar talker benefit only emerged for the words in clean and only for the listeners with a higher lexical proficiency which contradicts the time-course hypothesis.

The final research question concerned the role of proficiency of listeners in the emergence of the talker familiarity effects. We measured the lexical proficiency of the participants in our study. An effect of lexical proficiency was observed for the reaction times in the Old/New Task in which listeners with a higher lexical proficiency reacted faster to items produced by familiar talkers than to items produced by unfamiliar talkers when the items were in clean with the effect going in the same direction for the items in noise. Arguably, listeners with a higher lexical proficiency are more sensitive to the acoustic patterns in the speech signal and, as a consequence, better at discriminating old and new speakers. A similar explanation was offered by Trofimovich (2008) who suggested that more experienced non-native listeners are likely to be better at encoding context-specific phonological information from non-native words, and Levi et al. (2011) who underlined the importance of the ability of listeners to make a connection between talker-specific acoustic and language-specific phonetic and lexical information in the speech signal for the talker familiarity benefit to emerge.

The demonstration of the importance of speaker-related information in speech processing in the '90s (Mullennix et al., 1989; Nygaard et al., 1994; Nygaard & Pisoni, 1998; Palmeri et al., 1993) challenged abstractionist models of speech perception which discarded indexical information as irrelevant. A new exemplar-based account of speech perception was offered (Goldinger, 1998), postulating that each new occurrence of a word is stored as a unique memory trace containing all the information about this word in all its detail. Word recognition then entails the comparison of current input to all stored traces. The results of (more)

recent studies (Jesse et al., 2007; Kittredge et al., 2006; Lee & Zhang, 2015), including the present experiment, seem to suggest that indexical information is not saved as an integral part of the lexical representations but rather is stored in episodic memory.

Different studies (see, e.g., McLennan & Luce, 2005; Cutler, 2010) expressed the need for a hybrid model of speech perception activating and exploiting both abstract representations and more specific form-based representations. Several attempts have been made in formulating such a hybrid theory (e.g.Cutler, 2010; Goldinger, 2007; Kleinschmidt & Jaeger, 2015; Luce et al., 2003; McQueen, Cutler, & Norris, 2006). For instance, theories implying Bayesian inference argue that listeners make and update predictions about the speech signal based on the available evidence (Kleinschmidt & Jaeger, 2015; Norris & McQueen, 2008; Norris, McQueen, & Cutler, 2016). In this framework (e.g., Kleinschmidt & Jaeger, 2015), voice learning and the talker familiarity benefit can be explained by listeners creating a talker-specific generative model on the basis of talker-specific mappings of acoustic cues to phonetic categories. Listeners are able to recognize a familiar situation (familiar talker) and take advantage of this familiarity. At the same time, theories of this type imply that each successive input is used to update the belief of the listeners about the likelihood of a certain event occurring (Pufahl & Samuel, 2014), which could theoretically mean larger effects of talker-specific information for talkers to whom the listeners had more exposure. This is however not what we observe in our word recognition experiment with non-native listeners. We did not find an additional familiarity advantage despite the listeners having had extensive training on the voice of the talker. The failure to observe a talker familiarity benefit in word recognition however is in line with the theory put forward by Cutler et al. (2010). They suggested that the human spoken-word recognition system consists of abstract pre-lexical and lexical representations combined with an episodic memory system, where indexical information is stored, which is distinct from the mental lexicon but linked either to a linguistically abstract lexical or prelexical level. However, the circumstances under which each type of representation (abstract or voice-specific) are used during speech comprehension should be further investigated, comparing different groups of listeners (natives and non-natives), different voice conditions (same, familiar, or unfamiliar), and different tasks.

The present study demonstrated that familiarity with a talker's voice facilitates non-native speech processing but not non-native spoken-word

recognition. The effect of talker familiarity on speech processing was higher for listeners with a higher lexical proficiency, and increased when participants became more familiar with the voice of the speaker. Listening conditions influenced the emergence of the talker familiarity benefit, but the pattern of the effect differed depending on whether accuracy or reaction times were analyzed.

CHAPTER 7

Discussion and conclusions

The aim of the present thesis was to study the role of nativeness and the presence of noise on dealing with the variation in the speech signal introduced by variability within and between speakers. Previous studies revealed that suboptimal lexical and phonological knowledge, characteristic of non-native listeners, and suboptimal listening conditions due to the presence of background noise result in a decrease of recognition performance and an increase in processing cost for speech comprehension (e.g., Brouwer & Bradlow, 2011, 2016; Weber & Cutler, 2004; see for an overview Garcia Lecumberri et al., 2010). This is due to an impediment of the processibility and reliability of lexical, phonological and acoustic information in the speech signal, which in turn has repercussions for the interpretation of intra- and inter-speaker variability.

In this thesis, we specifically focused on perceptual learning as a mechanism to deal with the variability in the speech signal, namely, lexically-guided perceptual learning and the perceptual learning of voices. Three main research questions were addressed:

1. How does lexically-guided perceptual learning function in native and non-native listening in both clean and noisy listening conditions?

2. What factors influence perceptual learning of voices in non-native listening?

3. Does perceptual learning of voices facilitate speech comprehension in non-native listening in both clean and noisy listening conditions?

Chapters 2, 3 and 4 investigated questions related to the way lexically-guided perceptual learning functions in native and non-native listening in clean and noisy listening conditions, addressing the first research question. More specifically, Chapter 2 investigated whether lexically-guided perceptual learning occurs in a non-native language, Chapter 3 examined the effect of noise on lexically-guided perceptual learning in both native and non-native listening, and Chapter 4 considered the time course of lexical retuning in native listening. Perceptual learning of voices in a non-native language was investigated in Chapters 5 and 6. Chapter 5 dealt with the factors influencing perceptual learning of voices, as formulated in the second research question. Chapter 6 addressed the third research question about whether this perceptual learning facilitates speech comprehension in non-native listening in both clean and noisy listening conditions.

The remainder of this chapter summarizes the results obtained in the experiments presented in Chapters 2 to 6 in Sections 7.1 and 7.2. Subsequently, the similarities and differences regarding perceptual learning in native and non-native speech processing in clean and in noisy listening condition are discussed in Section 7.3. The findings in this thesis are then related to current theories of spoken word recognition in Section 7.4. The thesis ends with an outlook on possible directions for future research in Section 7.5 and conclusions in Section 7.6.

7.1 Adaptation to a talker's pronunciation

The aim of Chapter 2 was to investigate whether retuning of non-native phonetic categories is possible as a result of exposure to an ambiguous sound in a non-native language. Native British English and Dutch non-native listeners of English were exposed to an ambiguous sound halfway between /ɹ/ and /l/ which either substituted all /ɹ/ sounds or all /l/ sounds in an exposure short story. Importantly, the British English /ɹ/ does not occur as such in Dutch and thus constitutes a non-native phonetic category. Retuning was investigated in a subsequent phonetic

categorization task where participants had to categorize ambiguous items from the /ɹ/ to /l/ continuum as either containing /ɹ/ or /l/.

Two major findings emerged from this study. Firstly, ambiguous non-native sounds were found to induce lexically-guided perceptual learning of non-native phonetic categories, thus extending earlier results showing that ambiguous sounds in a non-native language can induce lexically-guided perceptual learning of native phonetic categories (Reinisch et al., 2013). Similar to the native perceptual system (Cutler, 2012), the non-native perceptual system is flexible and non-native phonetic category boundaries can be retuned. Secondly, perceptual learning was only observed for those listeners who had been exposed to the ambiguous sound in the words with /ɹ/. This retuning for only one sound in the sound pair suggests an asymmetry in lexically-guided perceptual learning. This asymmetry has previously been observed for the sounds /s/-/f/ (Eisner & McQueen, 2006; Norris et al., 2003; Zhang & Samuel, 2014). We postulated that this asymmetry is caused by a difference in the natural acoustic variation of the sounds: sounds that are inherently more acoustically variable are more easily retuned since listeners are used to hearing this variation and adapting these particular categories. There might be an upper limit to adaptation processes though. Kataoka and Koo (2017) found lexical retuning for an ambiguous /i/ but not for the more variable /u/. They hypothesized that adaptation leading to a (potential) overlap of the to-be-adapted phonetic category with other phonetic categories could block retuning of phonetic category boundaries.

Chapter 3 investigated the impact of the presence of background noise on the emergence of the lexically-guided perceptual learning effect in native and non-native listening. To that end, intermittent noise was added to the exposure story that was used in the experiment described in Chapter 2. A (new, separate from those tested in Chapter 2) group of native English listeners showed lexically-guided perceptual learning despite the presence of intermittent noise in the short story. The non-native Dutch listeners, however, did not show phonetic category retuning. These differences in results between the clean and noisy listening conditions for the non-native listeners could not be explained by a lower lexical proficiency or lower comprehension of the short story in the noise condition in comparison to the clean condition. Rather, the absence of learning by the non-native listeners when exposed to the short story in noise is argued to be connected to the larger effect of the presence of noise on the lexical competition process in non-native listening compared to that in

native listening. As described in Chapter 1, more words compete for activation in non-native compared to native listening due to imperfect sound perception and the subsequent spurious activation of words from both the native and the non-native language of the listeners (e.g., Broersma, 2012; Weber & Cutler, 2004). The presence of background noise increases this competition (e.g., Brouwer & Bradlow, 2011; Scharenborg et al., 2017). Moreover, when accustomed to degraded input, listeners might keep lexical competitors longer in memory (Farris-Trimble et al., 2014). The observed native versus non-native difference in perceptual learning in the presence of background noise is then postulated to be due to a slowing down of the recognition of the critical word due to the increase in the number of competitors simultaneously competing for recognition to the extent that the crucial disambiguating lexical information becomes available too late for lexically-guided perceptual learning to occur. The explanation that the necessary lexical information might be unavailable at the crucial point in time seems to line up with the hypothesis put forward by Jesse and McQueen (2011) that for lexically-guided perceptual learning to occur, disambiguating information for the ambiguous sound should be available early and reliably enough. In their study, retuning was blocked in native listening when ambiguous sounds occurred at the beginning of the words, while in our case the word containing the ambiguous sound was arguably not interpreted quickly enough, and consequently the ambiguous sound was recognized too late.

The results of the experiment on lexically-guided perceptual learning in native and non-native listening in the presence of intermittent background noise extend the findings by Zhang and Samuel (2014). They demonstrated an effect of noise on lexically-guided perceptual learning in native listening to the extent that category retuning was blocked when the whole stimulus was masked with noise except for the target sounds. Our study investigated the effect of intermittent noise (as opposed to noise masking the whole stimulus) not only in native but, as the first to do so, also in non-native listening. In addition, while Zhang and Samuel (2014) hypothesized that noise interferes with the reliability of the ambiguous sound which in turn interferes with category retuning, we postulate that intermittent noise does not reduce the reliability of the variability of the ambiguous sound but rather interferes with competition and word recognition processes so that the critical ambiguous sound is not disambiguated quickly enough. Taking the results of both studies together, we can draw the conclusion that the influence of noise

on lexically-guided perceptual learning is gradual and depends both on the amount of noise (intermittent or covering the whole stimuli) and the nativeness of the listeners (native or non-native).

Chapter 4 studies in how far items containing ambiguous sounds are perceived and processed as real words and whether adaptation to the ambiguous sound tends to equalize the processing of these items and their natural counterparts. In this study, native Dutch listeners were exposed to an ambiguous sound between /s/ and /f/ in a lexical decision task in Dutch. In the lexical decision task, items with the ambiguous sound were followed by semantically related words. Listeners adapted to the ambiguous sound as shown by a retuning of the phonetic category in a subsequent phonetic categorization task. Analysis of the items with the ambiguous sounds showed that these manipulated words were accepted slower and less often as real words compared to their natural counterparts, confirming that the presence of an ambiguous sound makes processing and recognition of these words different from their counterparts containing a natural sound (Scharenborg & Janse, 2013; Schuhmann, 2016). Semantically related words following the words with an ambiguous or a natural sound did not demonstrate such a difference between "natural" and manipulated words, suggesting that although the presence of an ambiguous sound slows down building up the activation of a word, the spreading of its activation to semantically related words occurred without delay. Similar to Poellmann et al. (2011), we found that processing eventually becomes more natural-like and adaptation to an ambiguous sound occurs in a stepwise manner. Finally, similar to Scharenborg and Janse (2013), Dutch native listeners who accepted more items as real words during the exposure phase, demonstrated more retuning, confirming the importance of recognizing the manipulated word as a real word for lexically-guided perceptual learning to occur (Norris et al., 2003).

To summarize, Chapters 2 to 4 showed that both native and non-native listeners can adapt to ambiguous pronunciations through retuning of their phonetic category boundaries. For lexically-guided perceptual learning to occur, it is important that the ambiguity is resolved early and reliably enough, a process which is arguably disrupted in the presence of background noise for non-native listeners. For native listeners it was again shown that this adaptation occurs fast and in a stepwise manner, and that it is stronger for those listeners who accept more words with an ambiguous sounds as real words. Whereas the time course of lexically-

guided perceptual learning and processing and recognition of items with ambiguous sounds has not been studied for non-native listeners, it is likely that it occurs in a similar manner to that of the native listeners studied in Chapter 4.

7.2 Adaptation to a talker's voice

Chapter 5 investigated the role of speaker-, listener- and stimulus-related factors on learning to recognize new voices speaking in a non-native language. Non-native listeners of Dutch learned to recognize four British English speakers in the course of four days. Each listener was exposed to a unique combination of four speakers selected from a pool of twelve male native English speakers. The listeners successfully learned the four voices relying on speaker-specific acoustics, such as fundamental frequency. Moreover, words containing more sounds carrying speaker-specific information were shown to be more beneficial for voice learning than words containing fewer of such sounds. Neither lexical frequency of the words nor lexical proficiency of the listeners were shown to influence listeners' learning progress. Two possible explanations were offered for the latter finding. First, it is possible that lexical information is not required for successful voice learning, which is in line with studies showing that listeners can identify talkers even with unintelligible linguistic content in time-reversed speech (Bricker & Pruzansky, 1966; Sheffert et al., 2002) or in a completely unfamiliar language (Winters et al., 2013). Second, other studies suggest that voice learning is not language independent, but that phonological information in combination with acoustic characteristics of voices rather than lexical information guides voice learning (Perrachione et al., 2011; Zarate et al., 2015). Although the usage of phonological information in voice learning was not explicitly tested in this study, arguably, the non-native listeners in Chapter 5 were familiar with the language of testing, meaning that they can rely on both language independent, as well as language specific phonological information to learn voices. Finally, listeners with a larger working memory capacity were shown to learn voices faster than listeners with a lower working memory capacity. Levi (2014) suggested that listeners with a larger working memory capacity might use a different strategy when learning to recognize new voices. Our results extend this finding, showing that listeners with a larger working memory capacity require less time to learn to recognize

voices.

After having established that Dutch non-native listeners are able to learn new English voices, Chapter 6 investigated whether familiarity with the talker facilitates non-native speech processing and word recognition in the clean and in the presence of background noise. To investigate this, the same listeners as in the voice learning experiment participated in two additional experiments. At the end of each day of the voice training, listeners had to perform a recognition memory task, which included both words already presented to the listeners and new words. Half of the old and half of the new words were produced by the talkers to whom the listeners were familiarized, and half of the words were produced by new, unfamiliar talkers. Additionally, half of the words in each of these four experimental sets were presented in speech-shaped noise and half in the clean. Listeners had to identify whether the word they were hearing had already been presented in the earlier tasks in the experiment or whether the word was new.

The results showed that listeners were more accurate when reacting to words produced by familiar talkers than to words produced by unfamiliar talkers. These results are in line with studies demonstrating that listeners can benefit from the match in indexical (talker-specific) information between the first and the second presentation of the word in both the native (Goh, 2005; Palmeri et al., 1993) and a non-native language (Winters et al., 2013). The non-native listeners thus demonstrated a familiar talker benefit in speech processing. Moreover, listeners with a higher lexical proficiency reacted faster to words by a familiar speaker than to words by an unfamiliar speaker, showing the role of language proficiency in the emergence of the familiar talker benefit. This finding is in line with previous findings by Trofimovich (2008) for the same talker advantage. In line with the time-course hypothesis (Luce et al., 2003), we found a significant difference in accuracy for the words spoken by familiar and unfamiliar talkers only for the words in noise. However, the reaction time analyses only showed a talker-familiarity effect for the words in clean. Thus, non-native listeners were shown to be able to store and use talker-specific information, where the advantage of hearing words by familiar talkers was dependent on the proficiency of the listeners and the listening condition.

To study the effect of talker familiarity on the word recognition process, a word recognition task was introduced to the listeners prior to the first voice training and at the end of the last voice training day. For one

group of listeners, the word recognition task was conducted with a voice on which they would be trained for four days (familiar talker condition), while the other group heard words spoken by a speaker on which they would not be trained (unfamiliar talker condition). Non-native listeners' progress from the first to the last experimental day was measured. Contrary to earlier studies, which found a familiar voice benefit in native word recognition (Nygaard & Pisoni, 1998; Nygaard et al., 1994), no difference in the increase in performance from the first to the last experimental day was observed for the non-native listeners in the familiar and the unfamiliar talker conditions. There are several possible explanations for this difference in the emergence of a talker familiarity benefit in word recognition and speech processing. One explanation is that non-native listeners were able to familiarize themselves with the voice of the talker already after the first exposure in the word recognition task on Day 1 leaving (very) little additional benefit of explicit voice training, an explanation similar to what was suggested by Newman and Evers (2007). Another explanation is related to the different nature of the tasks and how indexical information is stored in listeners' memory. It is hypothesized that talker-specific information is stored not as part of the lexical representation of a word but rather in an episodic memory system separate from the mental lexicon (Cutler, 2010; Cutler et al., 2010). Episodic memory is accessed during a recognition memory task but not during a word recognition task. Listeners thus benefit from the match in indexical information between the first and the second presentation of a word in a recognition memory task giving rise to a familiar talker effect in speech processing but not during word recognition.

To summarize, Chapters 5 and 6 showed that non-native listeners store talker-specific information in their memory and are able to benefit from familiarity with the voice of the talker, similar to what has been observed previously for native listeners. Non-native listeners seem to rely on both talker-specific, acoustic and language-specific, phonological information in voice learning, while talker-specific information (familiarity) assists listeners in speech processing. Talker-specific and language-specific information in the speech signal seem to interact in speech perception (see also, e.g., Nygaard & Pisoni, 1998; Levi et al., 2011; Winters et al., 2008). While lexical proficiency was not shown to influence voice learning itself, it was important for the emergence of talker familiarity effects in speech processing: only listeners with a higher lexical proficiency demonstrated a talker familiarity benefit. This finding is in line with Levi

et al. (2011), who showed no differences in voice learning performance for native listeners and listeners unfamiliar with the language of testing, but did observe a talker familiarity benefit only for the native listeners. Finally, we have demonstrated that the talker familiarity benefit depends on the nature of the task and the listening conditions.

7.3 Native and non-native speech perception

The present thesis revealed important similarities between native and non-native speech perception, such as the non-native perceptual system's flexibility in the face of speaker-related idiosyncrasies (Chapters 2 and 3), the ability to adapt to previously unfamiliar talkers while relying on acoustic, phonetic and possibly phonological information in the signal to do so (Chapter 5), and benefitting from familiarity with a talker in speech processing (Chapter 6). Non-native listeners, like native listeners, use the mechanism of perceptual learning to adapt to variability due to within and between speaker variation and the presence of background noise. The non-native listener's perceptual system dynamically learns and adjusts as a function of the received input.

Two essential characteristics of both the native and non-native perceptual system can be derived on the basis of these results: adaptive phonetic flexibility and the exploitation of talker-specific knowledge. Flexibility is revealed in the ability of listeners to adapt their phonetic category boundaries. Talker-specificity reveals itself in the ability of listeners to learn to recognize previously unfamiliar talkers and benefit from this knowledge in later speech perception. Both flexibility and the use of talker-specific information were shown to be dependent on the listening situation for the non-native listeners in the present study, while similar results were found in previous studies for native listeners (McLennan & Luce, 2005; Eisner & McQueen, 2006; Zhang & Samuel, 2014). Non-native listeners in Chapter 2 adapted their phonetic category boundary for a more acoustically variable /ɹ/, but not for a less acoustically variable /l/, and in Chapter 6 the reliance on talker-specific information depended on the presence or absence of background noise. These two observations are in line with the ideal adapter framework introduced by Kleinschmidt and Jaeger (2015), in which listeners use the cues available in the speech signal depending on their prior beliefs about reliability and informativity of these cues. For instance, according to this framework,

the listener's prior beliefs about a category (in our case /l/ or /ɹ/) constitute a prediction about the distribution of this category in the future. If the category is highly variable, listeners are more prepared to shift their beliefs about the category boundaries.

Despite the observed similarities in how native and non-native listeners deal with the variability in the speech signal, there are also differences. Firstly, variability introduced by intermittent noise has more repercussions for non-native than native listeners. While native listeners can flexibly adapt their phonetic category boundaries even when there is intermittent noise in the signal, lexically-guided adaptation is blocked in non-native listening. However, we hypothesize that this difference is not due to a difference in the effect of noise on the perceptual learning process, but rather due to a difference in the slowing down of the speech comprehension process. Noise affects native and non-native speech perception in similar ways (Scharenborg et al., 2017), i.e., it slows down the speech recognition process (Brouwer & Bradlow, 2011), and decreases listeners certainty in what words they are hearing (McQueen & Huettig, 2012). In the already suboptimal non-native listening situation where more words compete for activation due to the imperfect sound perception and spuriously activated words from both the native language of the listeners and the non-native language, the activation of a word with an ambiguous sound is hypothesized to be slowed down to the extent that the ambiguous sound is not disambiguated quickly enough for lexically-guided perceptual learning to occur.

We should bear in mind another important difference between native and non-native speech comprehension and that is the impoverished lexical knowledge of non-native compared to native listeners, which could negatively influence the recognition of words with an ambiguous sound and the use of acoustic-phonetic cues to process talker-related information. The non-native listeners participating in the studies described in the present thesis had on average an upper-intermediate proficiency in English. Their lexical knowledge was high enough to exploit the available lexical information to disambiguate the manipulated sounds and to guide the retuning of phonetic category boundaries in clean listening condition in Chapters 2 and 3. However, lexical proficiency was shown to influence the familiar talker benefit in Chapter 6. Only listeners with a higher proficiency showed a talker familiarity benefit in speech processing, which could be connected to the increased ability of listeners with a higher lexical proficiency in the non-native language to pick up

relevant acoustic and phonological information from the signal (Trofimovich, 2008), and therefore differentiate familiar and unfamiliar talkers better. Interestingly, this contradicts the idea put forward by Creel and Tumlin (2011), who suggested that the facilitatory effect of an acoustic match between the first and the second presentation of a word should be the strongest for listeners not experienced in the language. If listeners with a higher lexical proficiency are indeed more sensitive to phonetic detail in spoken words, as Trofimovich (2008) suggested, they should in principle also be better at learning to recognize and differentiate voices in a non-native language. This difference, however, was not observed in Chapter 5, where lexical proficiency was not shown to modulate voice learning performance of the listeners. Testing this hypothesis probably requires listeners with a much lower lexical proficiency or words with a lower frequency.

7.4 Perceptual learning in native and non-native listening in the perspective of current theories of spoken word recognition

There is an ongoing debate between the proponents of abstractionist and episodic models of spoken word recognition (see also Chapter 1). In line with McQueen, Cutler, & Norris (2006), we observed generalization of lexically-guided perceptual learning in both native and non-native listening to words not present in the exposure, which implies the need for abstract representations at the prelexical and lexical levels of processing (Cutler, 2010; Cutler et al., 2010; Cutler, 2017). Furthermore, we found no interaction of lexical and indexical information in the speech signal during voice learning. Lexical knowledge or lexical frequency of the items used to train the listeners to recognize previously unfamiliar voices were not shown to play a role in non-native voice learning. This is in line with multiple studies showing that it is not lexical, but rather phonological information that guides voice learning (Perrachione et al., 2011; Perrachione & Wong, 2007; Zarate et al., 2015). The benefit of knowledge of the phonology of a language rather than lexical knowledge in voice learning suggests the existence of separate processing levels, and an abstract representation of phonological knowledge (Cutler, 2017).

Note, however, that the usage of phonological information in voice learning means that linguistic and indexical information in the speech

signal interact during speech comprehension, which is in line with episodic theories of lexical access (Pisoni & Levi, 2007). At the same time, Chapter 6 provides evidence that listeners store indexical talker-related information in their memory, and that this information facilitates processing. This finding is in line with episodic theories of lexical access (Goldinger, 1996, 1998). Importantly, however, the facilitatory effect of talker familiarity was modulated by the presence of noise and the nature of the task, suggesting that talker-specific information is not always accessed during speech comprehension. A comprehensive theory of human speech comprehension should be able to account for both types of findings: those pointing at the role of abstract information in spoken-word recognition and those underlying the importance of indexical information.

Several attempts have been made to formulate such a hybrid theory (e.g., Cutler, 2010; Goldinger, 2007; Kleinschmidt & Jaeger, 2015; Luce et al., 2003; McQueen, Cutler, & Norris, 2006). The proposals differ in their explanation of where abstract and voice-specific information is stored and when these types of information are used during speech comprehension. We will discuss our results from the point of view of the time-course hypothesis (Luce et al., 2003), weak abstractionist theories suggesting storage of talker-specific detail in the episodic memory (e.g., Cutler et al., 2010) and theories implying Bayesian inference in speech processing (e.g., Kleinschmidt & Jaeger, 2015; Norris & McQueen, 2008; Norris et al., 2016).

Luce et al. (2003) suggested that abstract representations initially dominate processing, while indexical information affects processing later. In Chapter 6, listening conditions were shown to influence the emergence of talker familiarity effects on listeners' performance in an Old/New task. These effects were in line with the time-course hypothesis. At the same time, when reaction times were analyzed, the talker familiarity benefit was only observed for words in the clean, contradicting the time-course hypothesis. The results presented in this thesis can therefore not provide a definite answer to the question about the time course of indexical effects and the role of noise in the emergence of talker familiarity advantage.

An interesting explanation for the interaction between episodic and abstract information is offered by theories implying Bayesian inference in speech processing, which argue that listeners make and update predictions about the speech signal based on the available evidence (Kleinschmidt & Jaeger, 2015; Norris & McQueen, 2008; Norris et al., 2016).

In this framework (e.g., Kleinschmidt & Jaeger, 2015), perceptual learning can be explained by listeners creating a talker-specific generative model on the basis of talker-specific mappings of acoustic cues to phonetic categories. Moreover, the relevance of previous listening experience (a phonetic category, talker, specific talker-groups) depends on the familiarity with the talker and the listening situation. When encountering a novel situation, listeners are able to rapidly adapt, generalize this adaptation to similar situations, and, at the same time, recognize a familiar situation again and take advantage of this familiarity. Theories of this type, however, imply that each successive input is used to update the belief of the listeners about the likelihood of a certain event occurring (Pufahl & Samuel, 2014), which could theoretically mean larger effects of talker-specific information for talkers to whom the listeners had more exposure. This is not what we observe in Chapter 6. No additional benefit of larger familiarity with the voice of the talker was found in the word recognition task.

Cutler et al. (2010) put forward the idea that the human spoken-word recognition system consists of abstract pre-lexical and lexical representations combined with an episodic memory system, which is distinct from the mental lexicon but linked either to a linguistically abstract lexical or prelexical level. According to this theory, indexical information is not stored in the mental lexicon as a part of the lexical representation of a word neither is it stored at the pre-lexical level but rather it is stored in the listeners' episodic memory. Our findings in this thesis seem to support this idea, since we have observed generalization of lexically-guided perceptual learning and showed the storage of indexical information which manifested itself in a recognition memory task, but not in a word recognition task requiring lexical access. However, we should bear in mind that the word recognition task used in Chapter 6 used the same words in the pre- and post-test. We can therefore not fully ignore the possibility that indexical information was stored together with the lexical representation with the first presentation and is accessed during speech comprehension. However, we believe this explanation to be unlikely since a number of recent studies showed that indexical information is not normally accessed in the tasks requiring lexical access (Kittredge et al., 2006; Lee & Zhang, 2015), and talker-specific effects can emerge even when not the same words but the same phonemes (Jesse et al., 2007) or even non-words are used in the recognition memory task (Winters et al., 2013).

7.5 Directions for future research

While we have found both similarities and differences in perceptual learning in native and non-native listeners (see 7.3), not all the studies in the present thesis compared native and non-native listeners. Chapter 4 only addressed the time course of lexically-guided perceptual learning in native listening. Follow-up research should shed light onto the question of the time course of lexically-guided perceptual learning in non-native listening. Potentially, non-native listeners are slower in adapting to ambiguous sounds than native listeners, which might in turn affect the processing and recognition of the semantically related words following the items with ambiguous sounds, contrary to what was observed for the native listeners. Another open question concerns the talker familiarity benefit described in Chapter 6. It is possible that the lack of talker familiarity effects in non-native word recognition is not related to methodological differences with other familiar talker benefit studies conducted with native listeners (e.g., Nygaard & Pisoni, 1998), but rather to the native versus non-native difference. Testing native listeners using the same design can shed more light onto the role of (additional) familiarity with the speaker during word recognition. Moreover, the group of listeners tested in the present thesis had a relatively high proficiency in English. Testing a group of non-native listeners that is less uniform in their proficiency level might give more insight into the interaction of linguistic and indexical effects in speech perception and spoken word recognition, as well as pre-requisites for lexically-guided perceptual learning.

The talkers used in the studies on adaptation to multiple speakers in Chapters 5 and 6 differed in their learnability in the voice learning experiment as well as in their intelligibility in noise. In particular, in the word recognition task in Chapter 6, one speaker was more intelligible in noise than the other speaker, which could contribute to our failure to observe talker familiarity effects in this particular task. Schierloh and Hayes-Harb (2008) showed that for non-native listeners talker intelligibility plays a greater role than talker familiarity (note that all the tasks in their study were conducted in clean). While we have shown that talkers that were better to recognize based on their acoustic characteristics in a multinomial logistic regression analysis, were also better recognized by human listeners, the role of similarity and differences between voices in the talker familiarity effect was not addressed. Although there were attempts with mixed results to address this question in native listening

(e.g. Goh, 2005; Goldinger, 1996), no studies to our knowledge addressed the same issue in non-native listening. Further research could address the question regarding factors that influence learnability of a voice as well as whether higher learnability corresponds to a larger familiar talker benefit.

Adaptation to the voices of previously unfamiliar speakers and the emergence of talker-specific effects in non-native speech perception show that non-native listeners store and use indexical information, leading to the question of where exactly indexical information is stored (within or outside mental lexicon) and when it is used. The familiar-talker advantage observed in the recognition memory task and not in the word recognition task in the present study can theoretically be explained by the storage of indexical information in episodic memory. However, as suggested by Pufahl and Samuel (2014), to show that indexical information is not stored at the lexical level, one has to demonstrate indexical effects on the basis of non-words, as these necessarily will not involve lexical representations. Only one study to our knowledge indirectly addressed this question demonstrating a same talker advantage in a recognition memory task in a language unfamiliar to the listeners, suggesting that at least same talker advantage in speech perception is not word-specific (Winters et al., 2013). Furthermore, Jesse et al. (2007) showed the role of talker-specific information at the level of phonemes. While these findings showed that the prelexical processing level has access to talker-specific information, it does not necessarily mean that indexical information is not stored as part of a lexical representation. Further research is needed to address this question.

7.6 Conclusions

The studies described in the present thesis extend our understanding of listeners' flexibility in dealing with within and between-speaker variability by looking at the role of nativeness and the presence of background noise. We have demonstrated that non-native listeners, similar to native listeners, use perceptual learning to adapt to ambiguous pronunciations of one speaker and to the voices of multiple speakers in a non-native language. The presence of noise in the speech signal was shown to inter-

fere with the adaptation of non-native listeners to a talker's ambiguous pronunciations (but not for native listeners), and influenced the use of indexical information in non-native speech perception.

Appendix A
Clean exposure story used in Chapter 2 and 3

He opened the magazine, **immediately** [1] saw his own name, and began **wondering** how many fans had commented on his team's web page since Monday. He had been **ignoring** his phone, TV and the **Internet** since Monday evening, and wished the event to **quietly** fade out of his memory. His team had **happily** gone to an away game on Monday, but met an unexpected and **humiliating** defeat. It ended in a one-to-seven defeat against the **neighboring** city's team, **undoubtedly** thought to be the weakest of the two. The bookies gains on this one must have seemed **apparent** to anyone.

Nobody could **adequately** imagine that outcome: the team had **accumulated** wins and defeated opponents, attacking and defending with the **acquired** ease. It had **operated** as a machine does: it was fast and **accurate**. Magazines had been **admiring** him, speaking about his **inherent** gift as a coach, his **coherent** tactics, and his **ability** to change any team into one of the best **category**. But on Monday those outstanding **capabilities** vanished as if they had not once existed. The team showed a sudden **inability** to attack, **cooperate** and defend. He knew: he had to quit his post **immediately**. No **moderate** steps can be expected in this situation. It was so sad: his job had given him money, fame, and **mobility**.

Upon **entering** into the top-division competition, he hadn't expected to achieve anything. In an off-**camera** dialogue with a talk-show host, he even **openly** admitted it. But now the thought of having to join that

[1] Target words are in bold.

catalogue of coaches, each one queuing up to find a new coaching position, intimidated him. He expected no **equality** of chances: no famous team was going to invite him now as a coach. No one. That's enough, he thought. He had to face the situation and this **inequality** and pay no attention to **ignorant** fans. Act **independently** of what they might say. The exact moment he decided that mind-**wandering**, sitting and thinking about his devastating situation had no **utility**, somebody knocked at his window.

Appendix B
Noisified exposure story used in Chapter 3

He opened [1] the magazine, **immediately** saw his own name, and began **wondering** how many fans had commented on his team's web page since Monday. He had been **ignoring** his phone, TV and the **Internet** since Monday evening, and wished the event to **quietly** fade out of his memory. His team had **happily** gone to an away game on Monday, but met an unexpected and **humiliating** defeat. (pause). It ended in a one-to-seven defeat against the **neighboring** city's team, **undoubtedly** thought to be the weakest of the two. The bookies gains on this one must have seemed **apparent** to anyone.

Nobody could **adequately** imagine that outcome: the team had **accumulated** wins and defeated opponents, attacking and defending with the **acquired** ease. It had **operated** as a machine does: it was fast and **accurate**. Magazines had been **admiring** him, speaking about his **inherent** gift as a coach, his **coherent** tactics, and his **ability** to change any team into one of the best **category**. But on Monday those outstanding **capabilities** vanished as if they had not once existed. The team showed a sudden **inability** to attack, **cooperate** and defend. He knew: he had to quit his post **immediately**. (pause). No **moderate** steps can be expected in this situation. It was so sad: his job had given him money, fame, and **mobility**.

Upon **entering** into the top-division competition, he hadn't expected to achieve anything (pause). In an off-**camera** dialogue with a talk-show host, he even **openly** admitted it. But now the thought of having to join

[1] Underlined fragments are masked by noise.

that **catalogue** of coaches, each one queuing up to find a new coaching position, intimidated him. He expected no **equality** of chances: no famous team was going to invite him now as a coach. No one. That's enough, he thought. He had to face the situation and this **inequality** and pay no attention to **ignorant** fans. (pause). Act **independently** of what they might say. The exact moment he decided that mind **wandering**, sitting and thinking about his devastating situation had no **utility**, somebody knocked at his window.

Appendix C
Comprehension questions used in Chapter 2 and 3

1. What is the profession of the main character?
2. Which team was expected to win the last game?
3. How did his team perform in the last game?
4. What had he expected to achieve in a top-division competition?
5. Was he fired from his position or did he quit himself?

Appendix D
Prime-target pairs used in Chapter 4

F-final words			
Prime		Target	
one syllable primes			
boef	*rogue*	crimineel	*criminal*
braaf	*good*	ondeugend	*naughty*
gleuf	*slot*	opening	*opening*
lijf	*body*	lichaam	*body*
maf	*silly*	gek	*stupid*
plof	*flop*	geluid	*sound*
proef	*test*	eten	*food*
troef	*trump*	poker	*poker*

The table continues on the next page

F-final words			
Prime		Target	
two syllable primes			
actief	*active*	druk	*busy*
archief	*archive*	papier	*paper*
bankroof	*bank robbery*	geld	*money*
geloof	*faith*	religie	*religion*
giraf	*giraffe*	lang	*tall*
karaf	*carafe*	kruik	*jar*
kerkhof	*cemetery*	dood	*death*
witlof	*chicory*	groente	*vegetable*
three syllable primes			
achterneef	*second cousin*	oom	*uncle*
biograaf	*biographer*	boek	*book*
middenrif	*diaphragm*	luchtpijp	*trachea*
objectief	*objective*	neutraal	*neutral*

S-final words			
Prime		Target	
one syllable prmes			
bloes	*blouse*	hemd	*shirt*
buis	*tube*	riool	*drain*
grijs	*gray*	grauw	*gray*
kous	*stocking*	panty	*tights*
luis	*louse*	haar	*hair*
muis	*mouse*	kat	*cat*
poos	*while*	tijd	*time*
roes	*whirl*	alcohol	*alcohol*
two syllable primes			
atlas	*atlas*	kaart	*map*
iris	*iris*	oog	*eye*
kermis	*fair*	draaimolen	*carousel*
kompas	*compass*	richting	*direction*
matras	*mattress*	bed	*bed*
oppas	*babysit*	kinderen	*children*
paleis	*palace*	koning	*king*
tennis	*tennis*	racket	*racket*
three syllable primes			
grandioos	*magnificent*	enorm	*huge*
paradijs	*paradise*	hemel	*sky*
pindakaas	*peanut butter*	boterham	*sandwich*
rijbewijs	*driving licence*	auto	*car*

Bibliography

Abercombie, D. (1967). *Elements of general phonetics.* Aldine Pub. Company.

Abrams, D. A., Chen, T., Odriozola, P., Cheng, K. M., Baker, A. E., Padmanabhan, A., ... Menon, V. (2016). Neural circuits underlying mother's voice perception predict social communication abilities in children. *Proceedings of the National Academy of Sciences*, *113*(22), 6295–6300.

Alloway, T. P., Gathercole, S. E., Willis, C., & Adams, A.-M. (2004). A structural analysis of working memory and related cognitive skills in young children. *Journal of Experimental Child Psychology*, *87*(2), 85–106.

Amino, K., & Arai, T. (2008). Differential effects of the phonemes on identification of previously unknown speakers. *Journal of the Acoustical Society of America*, *123*(5), 3328–3335.

Amino, K., Arai, T., & Sugawara, T. (2007). Effects of the phonological contents on perceptual speaker identification. *Lecture Notes in Computer Science*, *4441*, 83–92.

Amino, K., Sugawara, T., & Arai, T. (2006). Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties. *Acoustical Science and Technology*, *27*(4), 233–235.

Andics, A. (2006). Distinguishing between prelexical levels in speech perception: an adaptation-fMRI study. *Nijmegen CNS*, *1*, 47–66.

Andics, A. (2013). *Who is talking? Behavioural and neural evidence for norm-based coding in voice identity learning* (Unpublished doctoral dissertation). Radboud University, Nijmegen.

Andics, A., McQueen, J. M., & Van Turennout, M. (2007). Phonetic content influences voice discriminability. In *16th international congress of phonetic sciences (ICPhS 2007)* (pp. 1829–1832). Pirrot.

Andruski, J. E., Blumstein, S. E., & Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition*, *52*(3), 163–187.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (release 2). Linguistic data consortium.*

Baddeley, A. D. (1986). *Working memory.* New York: Oxford University Press.

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation*, *8*, 47–89.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychological Research PRPF*, *74*(1), 110–120.

Ben-David, B. M., Chambers, C. G., Daneman, M., Pichora-Fuller, M. K., Reingold, E. M., & Schneider, B. A. (2011). Effects of aging and noise on real-time spoken word recognition: evidence from eye movements. *Journal of Speech, Language, and Hearing Research*, *54*(1), 243–262.

Bertelson, P., Vroomen, J., & De Gelder, B. (2003). Visual recalibration of auditory speech identification: a mcgurk aftereffect. *Psychological Science*, *14*(6), 592–597.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O. Bohn. & M. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13–34). Amsterdam/Philadelphia: John Benjamins Publising Company.

Boersma, P., & Weenink, D. (2009). *Praat: doing phonetics by computer (version 5.1. 05)[Computer program]. Retrieved May 1, 2009.*

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729.

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Attention, Perception, & Psychophysics*, *61*(2), 206–219.

Bradlow, A. R., & Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *The Journal of the Acoustical Society of America*, *106*(4), 2074–2085.

Bregman, M. R., & Creel, S. C. (2014). Gradient language dominance affects talker learning. *Cognition*, *130*(1), 85–95.

Bricker, P. D., & Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *The Journal of the Acoustical Society of America*, *40*(6), 1441–1449.

Broersma, M. (2012). Increased lexical activation and reduced competition in second-language listening. *Language and cognitive processes*, *27*(7-8), 1205–1224.

Brouwer, S., & Bradlow, A. R. (2011). The influence of noise on phonological competition during spoken word recognition. In *Proceedings of the international congress of phonetic sciences* (Vol. 2011, pp. 364–367).

Brouwer, S., & Bradlow, A. R. (2016). The temporal dynamics of spoken word recognition in adverse listening conditions. *Journal of Psycholinguistic Research*, *45*(5), 1151–1160.

Brouwer, S., Mitterer, H., & Huettig, F. (2012). Speech reductions change the dynamics of competition during spoken word recognition. *Language and Cognitive Processes*, *27*(4), 539–571.

Bull, R., Espy, K. A., & Wiebe, S. A. (2008). Short-term memory, working memory, and executive functioning in preschoolers: Longitudinal predictors of mathematical achievement at age 7 years. *Developmental Neuropsychology*, *33*(3), 205–228.

Chen, A., Gussenhoven, C., & Rietveld, T. (2004). Language-specificity in the perception of paralinguistic intonational meaning. *Language and Speech*, *47*(4), 311–349.

Chládková, K., Podlipskỳ, V. J., & Chionidou, A. (2017). Perceptual adaptation of vowels generalizes across the phonology and does not require local context. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(2), 414-427.

Clarke-Davidson, C. M., Luce, P. A., & Sawusch, J. R. (2008). Does perceptual learning in speech reflect changes in phonetic category representation or decision bias? *Attention, Perception, & Psychophysics*, *70*(4), 604–618.

Clopper, C. G., & Pisoni, D. B. (2004). Effects of talker variability on perceptual learning of dialects. *Language and Speech*, *47*(3), 207–238.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407.

Collins, B., & Mees, I. M. (1999). *The phonetics of English and Dutch*. Brill.

Cooke, M. (2009). Discovering consistent word confusions in noise. In *10th Annual Conference of the International Speech Communication Association* (pp. 1887–1890).

Creel, S. C., & Jimenez, S. R. (2012). Differences in talker recognition by preschoolers and adults. *Journal of Experimental Child Psychology*, *113*(4), 487–509.

Creel, S. C., & Tumlin, M. A. (2011). On-line acoustic and semantic interpretation of talker information. *Journal of Memory and Language*, *65*(3), 264–285.

Cutler, A. (2010). Abstraction-based efficiency in the lexicon. *Laboratory Phonology*, *1*(2), 301–318.

Cutler, A. (2012). Native listening: The flexibility dimension. *Dutch Journal of Applied Linguistics*, *1*(2), 169–187.

Cutler, A. (2017). Converging evidence for abstract phonological knowledge in speech processing. In *the 39th annual conference of the Cognitive Science Society (CogSci 2017)* (pp. 1447–1448).

Cutler, A., Eisner, F., McQueen, J. M., & Norris, D. (2010). How abstract phonemic categories are necessary for coping with speaker-related variation. *Laboratory Phonology*, *10*, 91–111.

Cutler, A., McQueen, J. M., Butterfield, S., & Norris, D. (2008). Prelexically-driven perceptual retuning of phoneme boudaries. In *the 9th Annual Conference of the International Speech Communication Association* (pp. 2056–2056).

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology*, *42*(4), 317–367.

Dart, S. N. (1991). Articulatory and acoustic properties of apical and laminal articulations. *UCLA Working Papers in Phonetics*, *79*, 1–155.

Dart, S. N. (1998). Comparing French and English coronal consonant articulation. *Journal of Phonetics*, *26*(1), 71–94.

De Deyne, S., & Storms, G. (2008). Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior Research Methods*, *40*(1), 198–205.

Drozdova, P., van Hout, R., & Scharenborg, O. (2014). Phoneme category retuning in a non-native language. In *the 15th Annual Conference of the International Speech Communication Association* (pp. 553–557).

Drozdova, P., van Hout, R., & Scharenborg, O. (2015). The effect of non-nativeness and background noise on lexical retuning. In *the 18th International Conference of the Phonetic Sciences.* Glasgow, UK: The Scottish Consortium for ICPhS 2015.

Drozdova, P., van Hout, R., & Scharenborg, O. (2016). Lexically-guided perceptual learning in non-native listening. *Bilingualism: Language and Cognition*, *19*(5), 914–920.

Eatock, J. P., & Mason, J. S. (1994). A quantitative assessment of the relative speaker discriminating properties of phonemes. In *Proceedings of icassp '94 (Adelaide, South Australia, April 1994)* (Vol. 1, pp. 1–133).

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *The Journal of the Acoustical Society of America*, *119*(4), 1950–1953.

Farris-Trimble, A., McMurray, B., Cigrand, N., & Tomblin, J. B. (2014). The process of spoken word recognition in the face of signal degradation. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(1), 308–327.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. *Speech Perception and Linguistic Experience: Issues in Cross-language Research*, *92*, 233–277.

Gallardo, L. F., Möller, S., & Wagner, M. (2015). Importance of intelligible phonemes for human speaker recognition in different channel bandwidths. In *the 16th Annual Conference of the International Speech Communication Association* (pp. 1047–1051).

Garcia Lecumberri, M. L., Cooke, M., & Cutler, A. (2010). Non-native speech perception in adverse conditions: A review. *Speech Communication*, *52*(11), 864–886.

Gibson, E. J. (1963). Perceptual learning. *Annual Review of Psychology*, *14*(1), 29–56.

Goggin, J. P., Thompson, C. P., Strube, G., & Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory & Cognition*, *19*(5), 448–458.

Goh, W. D. (2005). Talker variability and recognition memory: instance-specific and voice-specific effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 40–53.

Goldinger, S. D. (1996). Words and voices: episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1166–1183.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279.

Goldinger, S. D. (2007). A complementary-systems approach to abstract and episodic speech perception. In *Proceedings of the 16th international congress of phonetic sciences* (pp. 49–54).

Gordon, M., Barthmaier, P., & Sands, K. (2002). A cross-linguistic acoustic study of voiceless fricatives. *Journal of the International Phonetic Association*, *32*(2), 141–174.

Granena, G., & Long, M. (2013). *Cognitive aptitudes for second language learning and the LLAMA language aptitude test*. John Benjamins Publishing.

Hallé, P. A., Best, C. T., & Levitt, A. (1999). Phonetic vs. phonological influences on French listeners' perception of American English approximants. *Journal of Phonetics*, *27*(3), 281–306.

Hanulikova, A., & Ekstörm, J. (2017). Lexical adaptation to a novel accent in german: A comparison between german, swedish, and finnish listeners. In *the 18th Annual Conference of the International Speech Communication Association.*

Hintz, F., & Scharenborg, O. (2016). The effect of background noise on the activation of phonological and semantic information during spoken-word recognition. In *the 17th Annual Conference of the International Speech Communication Association* (pp. 2816–2820).

Jaeger, T. F. (2008). Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

Jesse, A., & McQueen, J. M. (2011). Positional effects in the lexical retuning of speech perception. *Psychonomic Bulletin & Review*, *18*(5), 943–950.

Jesse, A., McQueen, J. M., & Page, M. (2007). The locus of talker-specific effects in spoken-word recognition. In *16th International Congress of Phonetic Sciences (ICPhS 2007)* (pp. 1921–1924).

Jimenez, S. (2012). *The effect of language ability on talker identification* (Unpublished master's thesis). UC San Diego.

Johnson, E. K., Westrek, E., Nazzi, T., & Cutler, A. (2011). Infant ability to tell voices apart rests on language experience. *Developmental Science*, *14*(5), 1002–1011.

Kahneman, D. (1973). *Attention and effort* (Vol. 1063). Prentice-Hall Englewood Cliffs, NJ.

Kataoka, R., & Koo, H. (2017). Comparing malleability of phonetic category between [i] and [u]. *The Journal of the Acoustical Society of America*, *142*(1), 42–48.

Kawahara, H., Masuda-Katsuse, I., & De Cheveigne, A. (1999). Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, *27*(3), 187–207.

Kittredge, A., Davis, L., & Blumstein, S. E. (2006). Effects of nonlinguistic auditory variations on lexical processing in Broca's aphasics. *Brain and language*, *97*(1), 25–40.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203.

Köster, O., & Schiller, N. O. (1997). Different influences of the native language of a listener on speaker recognition. *Forensic Linguistics. The International Journal of Speech, Language and the Law*, *4*, 18–28.

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*, *107*(1), 54–81.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*(2), 141–178.

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, *13*(2), 262–268.

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*(1), 1–15.

Kraljic, T., & Samuel, A. G. (2011). Perceptual learning evidence for contextually-specific representations. *Cognition*, *121*(3), 459–465.

Kraljic, T., Samuel, A. G., & Brennan, S. E. (2008). First impressions and last resorts: How listeners adjust to speaker variability. *Psychological Science*, *19*(4), 332–338.

Latinus, M., & Belin, P. (2011). Anti-voice adaptation suggests prototype-based coding of voice identity. *Frontiers in Psychology*, *2*, 1–12.

Latinus, M., McAleer, P., Bestelmeyer, P. E., & Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Current Biology*, *23*(12), 1075–1080.

Laver, J. D. (1968). Voice quality and indexical information. *British Journal of Disorders of Communication*, *3*(1), 43–54.

Lee, C.-Y., & Zhang, Y. (2015). Processing speaker variability in repetition and semantic/associative priming. *Journal of Psycholinguistic Research*, *44*(3), 237–250.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, *44*(2), 325–343.

Levi, S. V. (2014). Individual differences in learning talker categories: The role of working memory. *Phonetica*, *71*(3), 201–226.

Levi, S. V. (2015). Talker familiarity and spoken word recognition in school-age children. *Journal of Child language*, *42*(4), 843–872.

Levi, S. V., & Schwartz, R. G. (2013). The development of language-specific and language-independent talker processing. *Journal of Speech, Language, and Hearing Research*, *56*(3), 913–925.

Levi, S. V., Winters, S., & Pisoni, D. B. (2008). A cross-language familiar talker advantage. *The Journal of the Acoustical Society of America*, *123*(5), 3331.

Levi, S. V., Winters, S. J., & Pisoni, D. B. (2011). Effects of cross-language voice training on speech perception: Whose familiar voices are more intelligible? *The Journal of the Acoustical Society of America*, *130*(6), 4053–4062.

Luce, P. A., & Lyons, E. A. (1998). Specificity of memory representations for spoken words. *Memory & Cognition*, *26*(4), 708–715.

Luce, P. A., McLennan, C. T., & Chance-Luce, J. (2003). Abstractness and specificity in spoken word recognition: Indexical and allophonic variability in long-term repetition priming. In J. Bowers & C. Marsolek (Eds.), *Rethinking Implicit Memory* (pp. 197–214). Oxford: Oxford University Press.

Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin*, *98*(1), 185–199.

Maibauer, A. M., Markis, T. A., Newell, J., & McLennan, C. T. (2014). Famous talker effects in spoken word recognition. *Attention, Perception, & Psychophysics*, *76*(1), 11–18.

Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, *8*(1), 1–71.

Mattys, S. L., Carroll, L. M., Li, C. K., & Chan, S. L. (2010). Effects of energetic and informational masking on speech segmentation by native and non-native speakers. *Speech Communication*, *52*(11), 887–899.

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7-8), 953–978.

Mattys, S. L., & Liss, J. M. (2008). On building models of spoken-word recognition: When there is as much to learn from natural "oddities" as artificial normality. *Attention, Perception, & Psychophysics*, *70*(7), 1235–1242.

Mayo, L. H., Florentine, M., & Buus, S. (1997). Age of second-language acquisition and perception of speech in noise. *Journal of Speech, Language, and Hearing Research*, *40*(3), 686–693.

McClelland, J. L., & Elman, J. L. (1986). The trace model of speech perception. *Cognitive Psychology*, *18*(1), 1–86.

McLennan, C. T. (2007). Challenges facing a complementary-systems approach to abstract and episodic speech perception. *Psychology Faculty Publications*, *19*.

McLennan, C. T., & González, J. (2012). Examining talker effects in the perception of native-and foreign-accented speech. *Attention, Perception, & Psychophysics*, *74*(5), 824–830.

McLennan, C. T., & Luce, P. A. (2005). Examining the time course of indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 306–321.

McQueen, J. M. (2005). Speech perception. In K. Lamberts & R. Goldstone (Eds.), *The Handbook of Cognition* (pp. 255–275). London: Sage Publications.

McQueen, J. M. (2007). Eight questions about spoken-word recognition. In M. G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 37–53). Oxford: Oxford University Press.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*(6), 1113–1126.

McQueen, J. M., & Huettig, F. (2012). Changing only the probability that spoken words will be distorted changes how they are recognized. *The Journal of the Acoustical Society of America*, *131*(1), 509–517.

McQueen, J. M., Norris, D., & Cutler, A. (2006). The dynamic nature of speech perception. *Language and Speech*, *49*(1), 101–112.

McQueen, J. M., Tyler, M. D., & Cutler, A. (2012). Lexical retuning of children's speech perception: Evidence for knowledge about words' component sounds. *Language Learning and Development*, *8*(4), 317–339.

Meador, D., Flege, J. E., & Mackay, I. R. (2000). Factors affecting the recognition of words in a second language. *Bilingualism: Language and Cognition*, *3*(1), 55–67.

Meara, P. (2005). *Llama language aptitude tests: The manual*. Swansea: Lognostics.

Mitterer, H., Chen, Y., & Zhou, X. (2011). Phonological abstraction in processing lexical-tone variation: Evidence from a learning paradigm. *Cognitive Science*, *35*(1), 184–197.

Mitterer, H., Scharenborg, O., & McQueen, J. M. (2013). Phonological abstraction without phonemes in speech perception. *Cognition*, *129*(2), 356–361.

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *The Journal of the Acoustical Society of America*, *85*(1), 365–378.

Neger, T. M., Rietveld, T., & Janse, E. (2014). Relationship between perceptual learning in speech and statistical learning in younger and older adults. *Frontiers in Human Neuroscience*, *8*.

Newman, R. S., & Evers, S. (2007). The effect of talker familiarity on stream segregation. *Journal of Phonetics*, *35*(1), 85–103.

Nijveld, A., ten Bosch, L., & Ernestus, M. (2015). Exemplar effects arise in a lexical decision task, but only under adverse listening conditions. In *the 18th International Congress of Phonetic Sciences (ICPhS 2015)*. Glasgow: University of Glasgow.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*(3), 189–234.

Norris, D., & McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395.

Norris, D., McQueen, J. M., & Cutler, A. (1995). Competition and segmentation in spoken-word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(5), 1209–1228.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238.

Norris, D., McQueen, J. M., & Cutler, A. (2016). Prediction, bayesian inference and feedback in speech recognition. *Language, Cognition and Neuroscience*, *31*(1), 4–18.

Nygaard, L. C. (2005). Perceptual integration of linguistic and nonlinguistic properties of speech. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 390–413). Blackwell Publishing Ltd.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Attention, Perception, & Psychophysics*, *60*(3), 355–376.

Nygaard, L. C., Sidaras, S. K., & Alexander, J. E. (2008). Time course of talker-specific learning in spoken word recognition. *The Journal of the Acoustical Society of America*, *124*(4), 2459–2459.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*(1), 42–46.

Owren, M. J., & Cardillo, G. C. (2006). The relative roles of vowels and consonants in discriminating talker identity versus word meaning. *The Journal of the Acoustical Society of America*, *119*(3), 1727–1739.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(2), 309–328.

Papcun, G., Kreiman, J., & Davis, A. (1989). Long-term memory for unfamiliar voices. *The Journal of the Acoustical Society of America*, *85*(2), 913–925.

Perrachione, T. K., Del Tufo, S. N., & Gabrieli, J. D. (2011). Human voice recognition depends on language ability. *Science*, *333*(6042), 595–595.

Perrachione, T. K., & Wong, P. C. (2007). Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex. *Neuropsychologia*, *45*(8), 1899–1910.

Pisoni, D. B., & Levi, S. V. (2007). Some observations on representations and representational specificity in speech perception and spoken word recognition. In M. G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 3–18). Oxford University Press Oxford.

Poellmann, K., McQueen, J. M., & Mitterer, H. (2011). The time course of perceptual learning. In *the 17th International Congress of Phonetic Sciences (ICPhS 2011)* (pp. 1618–1621). Pirrot.

Powell, M. J. (2009, June). *The BOBYQA algorithm for bound constrained optimization without derivatives* (Tech. Rep. No. NA2009/06). Cambridge, England: Centre for Mathematical Sciences, University of Cambridge.

Pufahl, A., & Samuel, A. G. (2014). How lexical is the lexicon? evidence for integrated auditory memory representations. *Cognitive Psychology*, *70*, 1–30.

Reinisch, E., & Holt, L. L. (2014). Lexically guided phonetic retuning of foreign-accented speech and its generalization. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(2), 539–555.

Reinisch, E., Weber, A., & Mitterer, H. (2013). Listeners retune phoneme categories across languages. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(1), 75–86.

Rogers, C. L., Lister, J. J., Febo, D. M., Besing, J. M., & Abrams, H. B. (2006). Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing. *Applied Psycholinguistics*, *27*(3), 465–485.

Rosenthal, E. N., Riccio, C. A., Gsanger, K. M., & Jarratt, K. P. (2006). Digit span components as predictors of attention problems and executive functioning in children. *Archives of Clinical Neuropsychology*, *21*(2), 131–139.

Ryalls, B. O., & Pisoni, D. B. (1997). The effect of talker variability on word recognition in preschool children. *Developmental Psychology*, *33*(3), 441–452.

Samuel, A. G. (2011). Speech perception. *Annual Review of Psychology*, *62*, 49–72.

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics*, *71*(6), 1207–1218.

Schacter, D. L., & Church, B. A. (1992). Auditory priming: implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(5), 915–930.

Scharenborg, O., Coumans, J. M., & van Hout, R. (2017). *The effect of background noise on the word activation process in non-native spoken-word recognition.* (to appear in Journal of Experimental Psychology: Learning, Memory, and Cognition)

Scharenborg, O., & Janse, E. (2013). Comparing lexically guided perceptual learning in younger and older listeners. *Attention, Perception, & Psychophysics*, *75*(3), 525–536.

Scharenborg, O., Kolkman, E., Kakouros, S., & Post, B. (2016). The effect of sentence accent on non-native speech perception in noise. In *the 17th Annual Conference of the International Speech Communication Association (Interspeech 2016)* (pp. 863–867).

Scharenborg, O., Mitterer, H., & McQueen, J. M. (2011). Perceptual learning of liquids. In *the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011)* (pp. 149–152).

Scharenborg, O., Weber, A., & Janse, E. (2015). The role of attentional abilities in lexically guided perceptual learning by older listeners. *Attention, Perception, & Psychophysics*, *77*(2), 493–507.

Schierloh, M., & Hayes-Harb, R. (2008). The contributions of talker familiarity and individual talker characteristics. *Die Unterrichtspraxis/Teaching German*, *41*(2), 171–185.

Schmidt, J., Scharenborg, O., & Janse, E. (2015). Semantic processing of spoken words under cognitive load in older listeners. In *the 18th International Congress of Phonetic Sciences (ICPhS 2015).* Glasgow: University of Glasgow.

Schuhmann, K. S. (2016). Cross-linguistic perceptual learning in advanced second language listeners. *Proceedings of the Linguistic Society of America*, *1*, 31–1.

Scobbie, J. M., Sebregts, K., & Stuart-Smith, J. (2009). Dutch rhotic allophony, coda weakening, and the phonetics-phonology interface. *QMU Speech Science Research Centre. Working Papers*, *WP-18*.

Sebregts, K. (2015). *The sociophonetics and phonology of dutch r* (Unpublished doctoral dissertation). Utrecht University.

Sheffert, S. M., Pisoni, D. B., Fellowes, J. M., & Remez, R. E. (2002). Learning to recognize talkers from natural, sinewave, and reversed speech samples. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(6), 1447–1469.

Sidtis, D., & Kreiman, J. (2012). In the beginning was the familiar voice: personally familiar voices in the evolutionary and contemporary biology of communication. *Integrative Psychological and Behavioral Science*, *46*(2), 146–159.

Snijders, T. A., B., & Bosker, R., J. (2012). *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. Sage.

Speciale, G., Ellis, N. C., & Bywater, T. (2004). Phonological sequence learning and short-term store capacity determine second language vocabulary acquisition. *Applied Psycholinguistics*, *25*(2), 293–321.

Theodore, R. M., Blumstein, S. E., & Luthra, S. (2015). Attention modulates specificity effects in spoken word recognition: Challenges to the time-course hypothesis. *Attention, Perception, & Psychophysics*, *77*(5), 1674–1684.

Torgesen, J. K., Rashotte, C. A., & Wagner, R. K. (1999). *Comprehensive test of phonological processing (CTOPP)*. Pro-ed Austin, TX.

Trofimovich, P. (2005). Spoken-word processing in native and second languages: An investigation of auditory word priming. *Applied Psycholinguistics*, *26*(4), 479–504.

Trofimovich, P. (2008). What do second language listeners know about spoken words? Effects of experience and attention in spoken word processing. *Journal of Psycholinguistic Research*, *37*(5), 309–329.

Tuft, S. E., McLennan, C. T., & Krestar, M. L. (2016). Hearing taboo words can result in early talker effects in word recognition for female

listeners. *The Quarterly Journal of Experimental Psychology*, *0*(0), 1-16.

Van de Velde, H., & van Hout, R. (1999). The pronunciation of (r) in standard dutch. *Linguistics in the Netherlands*, *16*(1), 177–188.

Van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A new and improved word frequency database for british english. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190.

Van Linden, S., & Vroomen, J. (2007). Recalibration of phonetic categories by lipread speech versus lexical information. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(6), 1483.

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language*, *50*(1), 1–25.

Weber, A., & Scharenborg, O. (2012). Models of spoken-word recognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*(3), 387–401.

Wechsler, D. (2008). Wechsler adult intelligence scale–fourth edition (WAIS–IV). *San Antonio, TX: NCS Pearson*, *22*, 498.

Whalen, D. H. (1991). Subcategorical phonetic mismatches and lexical access. *Attention, Perception, & Psychophysics*, *50*(4), 351–360.

White, K. S., Yee, E., Blumstein, S. E., & Morgan, J. L. (2013). Adults show less sensitivity to phonetic detail in unfamiliar words, too. *Journal of Memory and Language*, *68*(4), 362–378.

Winters, S., Levi, S. V., & Pisoni, D. B. (2008). Identification and discrimination of bilingual talkers across languages a. *The Journal of the Acoustical Society of America*, *123*(6), 4524–4538.

Winters, S., Lichtman, K., & Weber, S. (2013). The role of linguistic knowledge in the encoding of words and voices in memory. In *Second Language Research Forum, Ames, Iowa.*

Xie, X., & Myers, E. (2015). The impact of musical training and tone language experience on talker identification. *The Journal of the Acoustical Society of America*, *137*(1), 419–432.

Yonan, C. A., & Sommers, M. S. (2000). The effects of talker familiarity on spoken word identification in younger and older listeners. *Psychology and Aging*, *15*(1), 88–99.

Yovel, G., & Belin, P. (2013). A unified coding strategy for processing faces and voices. *Trends in Cognitive Sciences*, *17*(6), 263–271.

Zarate, J. M., Tian, X., Woods, K. J., & Poeppel, D. (2015). Multiple levels of linguistic and paralinguistic features contribute to voice recognition. *Scientific Reports*, *5*, 11475.

Zhang, X., & Samuel, A. G. (2014). Perceptual learning of speech under optimal and adverse conditions. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(1), 200–217.

Samenvatting in het Nederlands

Het spraaksignaal is zeer variabel. Luisteraars hebben van doen met een veelvoud aan sprekers, ieder met hun eigen unieke uitspraak veroorzaakt door factoren als dialect, accent, leeftijd, geslacht, grootte van het spraakkanaal of een spraakgebrek. Bovendien vindt communicatie vaak plaats in omstandigheden die verre van optimaal zijn, bijvoorbeeld door de aanwezigheid van achtergrondgeluid (ruis) of omdat er gecommuniceerd wordt in een taal die niet de moedertaal is van de luisteraars, waardoor het moeilijker is voor hen om de spraak van hun gesprekspartner te begrijpen. Hoe komen luisteraars tot de juiste interpretatie van het spraaksignaal ondanks deze variabiliteit en ondanks ongunstige luisteromstandigheden?

Volgens bestaande spraakherkenningstheorieën mappen luisteraars bij het herkennen van gesproken woorden het variabele signaal op reeds in het geheugen aanwezige representaties. De aard van deze representaties en de wijze van de verwerking van het signaal zijn nog steeds onderwerp van discussie. Abstractionistische theorieën stellen dat het spraaksignaal eerst wordt vergeleken met abstracte subwoordrepresentaties op het prelexicale verwerkingsniveau, waarna die subwoordrepresentaties die lijken op het spraaksignaal geactiveerd worden. Vervolgens activeren deze subwoordrepresentaties de lexicale representaties waarvan ze deel uitmaken. Dit gebeurt op het lexicale niveau. Episodische theorieën stellen daarentegen dat mensen stukken spraaksignaal gecodeerd met alle fonetische details, zoals bijvoorbeeld de stem van een spreker, in het geheugen opslaan en dat het binnenkomende spraaksignaal direct met deze reeds bestaande akoestische 'sporen' wordt vergeleken zonder dat er een prelexicaal niveau nodig is. Alle theorieën zijn het er evenwel over eens

dat bij het horen van een spraaksignaal meerdere lexicale kandidaten (abstractionistische theorieën) of sporen (episodische theorieën) gelijktijdig worden geactiveerd en strijden om herkenning. De winnende kandidaat is de kandidaat die het beste overeenkomt met het binnenkomende signaal.

Eerder onderzoek heeft laten zien dat moedertaalluisteraars zich snel kunnen aanpassen aan spreker-gerelateerde eigenaardigheden en dat ze de stem van onbekende sprekers leren herkennen door middel van een proces dat "perceptueel leren" heet. Luisteraars zijn in staat om deze aanpassing te generaliseren naar woorden die ze niet eerder hebben gehoord, wat suggereert dat er een bepaalde abstractie nodig is tijdens de spraakverwerking. Tegelijkertijd is gebleken dat moedertaalluisteraars profijt hebben van het bekend zijn met de stem van de spreker tijdens spraakherkenning wat aantoont dat spreker-gerelateerde informatie kan worden gebruikt door luisteraars tijdens het herkennen van spraak en dat deze dus in hun geheugen opgeslagen moet zijn. Het voordeel van het horen van een bekende stem is vooral groot als de luisteromstandigheden ongunstig zijn, hoewel eerder onderzoek suggereert dat perceptueel leren kan worden belemmerd als luisteraars moeite hebben om relevante akoestische, lexicale of fonologische informatie van het signaal op te pikken, zoals het geval zou kunnen zijn bij niet-moedertaalluisteraars. Luisteren in een niet-moedertaal kan moeilijk zijn vanwege de verschillen in klankinventaris tussen verschillende talen. Bovendien kan de niet-moedertaalluisteraar onvolledige lexicale kennis hebben van de niet-moedertaal. Het luisteren in een vreemde taal kan leiden tot problemen in het herkennen van de klanken in die niet-moedertaal, wat, zo blijkt uit de literatuur, leidt tot een toename in het aantal (onterecht) geactiveerde woorden tijdens luisteren in een niet-moedertaal en dus tot additionele lexicale concurrentie.

Dit proefschrift onderzoekt de effecten van ruis en het luisteren in een niet-moedertaal ('non-nativeness') op perceptueel leren van zowel de ambigue uitspraak van een spreker (dat proces wordt lexicaal-gestuurd perceptueel leren genoemd) als de stemmen van meerdere sprekers. Meer specifiek ga ik in op de volgende drie onderzoeksvragen: (1) Hoe functioneert lexicaal-gestuurd perceptueel leren in geval van luisteren in een moedertaal en in een niet-moedertaal taal en in luisteromstandigheden met en zonder ruis? (2) Welke factoren beïnvloeden het perceptueel leren van stemmen in een niet-moedertaal? (3) Faciliteert het perceptueel leren van stemmen spraakherkenning in een niet-moedertaal in luisterom-

standigheden met en zonder ruis?

In hoofdstuk 2 t/m 4 van dit proefschrift heb ik gekeken naar het perceptueel leren van een ambigue uitspraak in moedertaalluisteraars en niet-moedertaalluisteraars. Dit type perceptueel leren wordt lexicaal-gestuurd perceptueel leren genoemd omdat luisteraars lexicale informatie gebruiken om de ambiguïteit op te lossen. Eerder onderzoek heeft aangetoond dat lexicaal-gestuurd perceptueel leren zich manifesteert in een verschuiving van de grenzen van fonetische categorieën zodat bijvoorbeeld in het geval van een ambigue klank die het midden houdt tussen /s/ en /f/ de klank als een /s/ of /f/ wordt geïnterpreteerd afhankelijk van de lexicale context waarin de klank zich bevindt. In hoofdstuk 2 heb ik de vraag onderzocht of luisteraars in staat zijn om in een niet-moedertaal de grenzen van hun fonetische categorieën te verschuiven. Tijdens twee experimenten, die in dit hoofdstuk worden beschreven, luisterden Nederlandse niet-moedertaalluisteraars en Engelse moedertaal-luisteraars naar een kort verhaal in het Engels waarin alle /l/ dan wel /ɹ/-klanken waren vervangen door een ambigue klank die het midden hield tussen /l/ en /ɹ/. Een daaropvolgende fonetische categorisatietaak, waarbij luisteraars van items op een continuüm van *collect* naar *correct* en van *alive* naar *arrive* aan moesten geven of ze volgens hen een /l/ of /ɹ/ bevatten, liet een lexicaal-gestuurd perceptueel leereffect zien in zowel moedertaalluisteraars als niet-moedertaalluisteraars. Luisteraars die werden blootgesteld aan het verhaal waarin alle /ɹ/-klanken waren vervangen door de ambigue klank gaven meer /ɹ/-antwoorden in de fonetische categorisatietaak dan een groep luisteraars die werd blootgesteld aan een versie van het verhaal waarin alle /l/-klanken ambigu waren, en meer /ɹ/-antwoorden dan een derde groep, de baselinegroep, die niet werd blootgesteld aan het verhaal. Aangezien de intervocalische /ɹ/-klank geen deel uitmaakt van de klankeninventaris van het Nederlands, concludeer ik dat de Nederlandse luisteraars in staat waren om een niet-moedertaal klankcategorie (die dus is aangeleerd tijdens het leren van het Engels) aan te passen op basis van spraak in een niet-moedertaal. Deze resultaten laten een flexibiliteit van het perceptuele systeem zien die eerder alleen nog bij moedertaalluisteraars was aangetoond. Bovendien werd het bestaan van deze verschuiving aangetoond met behulp van woorden die niet waren gebruikt in het verhaal dat de luisteraars te horen kregen, hetgeen bewijs levert voor het bestaan van abstracte representaties op het prelexicale verwerkingsniveau.

In hoofdstuk 3 heb ik de grenzen van deze flexibiliteit verder onderzocht en heb ik onderzocht of af en toe aanwezige achtergrondruis interfereert met lexicaal-gestuurd perceptueel leren in geval van luisteren in een moedertaal en een niet-moedertaal. Eerdere studies hebben aangetoond dat de aanwezigheid van ruis in het spraaksignaal het aantal kandidaatwoorden dat strijdt om activatie verhoogt. Het gevolg is dat luisteraars minder zeker zijn van de woorden die ze horen, wat potentieel het herkenningsproces vertraagt. Om de rol van ruis in lexicaal-gestuurd perceptueel leren te onderzoeken kregen moedertaalluisteraars en niet-moedertaal-luisteraars de ambigue klank te horen in hetzelfde korte verhaal dat werd gebruikt voor de in hoofdstuk 2 gepresenteerde experimenten, met het verschil dat er nu korte periodes van achtergrondruis waren toegevoegd. Hoewel er nooit achtergrondgeluid werd geplaatst op de woorden met de ambigue klank en (op één uitzondering na) niet op de woorden vooraf-gaand aan en volgend op deze essentiële woorden, bleek de aanwezigheid van achtergrondgeluid het lexicaal-gestuurd perceptueel leren te verstoren bij de niet-moedertaalluisteraars, maar niet in dat van de moedertaal-luisteraars. De moedertaalluisteraars bleken perceptueel te leren in luister-omstandigheden met en zonder ruis. Het verschil voor de niet-moedertaal-luisteraars tussen het luisteren in omstandigheden met en zonder ruis kan niet worden toegeschreven aan verschillen in taalvaardigheid in het Engels (gemeten in lexicale kennis) of aan een slechter begrip van het verhaal door de niet-moedertaalluisteraars die naar het verhaal luisterden met ruis in vergelijking met de niet-moedertaalluisteraars die naar het verhaal luisterden zonder ruis. Mijn verklaring is dat niet-moedertaalluisteraars niet snel genoeg het woord met de doelklank herkennen om de ambigue klank op tijd te interpreteren als /ɹ/ of /l/ om lexicaal-gestuurd perceptueel leren te laten plaatsvinden. De reden voor deze vertraging is de activatie van additionele woorden in vergelijking met luisteromstandigheden zonder ruis. Eerder onderzoek heeft aangetoond dat wanneer de hele stimulus wordt gemaskeerd met ruis, lexicaal-gestuurd perceptueel leren ook wordt geblokkeerd in geval van luisteren in de moedertaal. In combinatie met mijn resultaten concludeer ik dat de invloed van ruis op lexicaal-gestuurd perceptueel leren geleidelijk is en afhankelijk is van zowel de locatie van de ruis (af en toe aanwezig of aanwezig over de hele stimulus) als van de 'nativeness' van de luisteraars.

Eerder onderzoek en de in hoofdstuk 2 en 3 beschreven studies hebben aangetoond dat moedertaalluisteraars en niet-moedertaalluisteraars in

staat zijn om de ambigue klank heel snel in hun klanksysteem op te nemen. Het experiment dat wordt beschreven in hoofdstuk 4 onderzocht de vraag hoe de woorden met de ambigue klank worden waargenomen en verwerkt door moedertaalluisteraars. In dit experiment werden Nederlandse moedertaalluisteraars tijdens een lexicale beslissingstaak inclusief een auditieve semantische primingstaak blootgesteld aan een ambigue klank die het midden hield tussen /s/ en /f/. De resultaten toonden aan dat de luisteraars langzamer waren in het accepteren van de woorden met deze ambigue klank dan echte woorden en ze minder vaak accepteerden als echte woorden dan dezelfde woorden die de natuurlijke klank bevatten. Toch weken acceptatie als een echt woord en de reactietijden voor semantisch verwante woorden die direct volgden op de woorden met de ambigue klank niet af van de woorden die direct volgden op de woorden met de natuurlijke /s/ of /f/. Deze bevindingen laten zien dat hoewel woorden met een ambigue klank anders worden verwerkt dan natuurlijke woorden, de verspreiding van hun activatie naar semantisch verwante woorden plaatsvindt zonder vertraging en niet lijkt te worden belemmerd. Interessant is dat aan het einde van de lexicale beslissingstaak het verschil tussen herkenning van de woorden met de ambigue klank en herkenning van dezelfde woorden met een natuurlijke klank is verdwenen, wat suggereert dat luisteraars hebben geleerd om de ambigue klank te interpreteren als een normaal geval van /s/ of /f/. Lexicaal-gestuurd perceptueel leren werd inderdaad aangetoond in een daaropvolgende fonetische categorisatietaak en de luisteraars die meer woorden accepteerden met de ambigue klank als echte woorden tijdens de lexicale beslissingstaak toonden ook een groter lexicaal-gestuurd perceptueel leereffect in vergelijking met de luisteraars die minder woorden als echte woorden accepteerden.

Hoofdstuk 5 en 6 zijn gewijd aan een andere vorm van perceptueel leren, namelijk perceptueel leren van stemmen, waardoor luisteraars bekend raken met de stemmen van tot dan toe onbekende sprekers. In hoofdstuk 5 heb ik gekeken naar drie groepen factoren die mogelijk van invloed kunnen zijn op het leren van stemmen in geval van luisteren in een niet-moedertaal: spreker-gerelateerde, luisteraar-gerelateerde en stimulus-gerelateerde factoren. Hoofdstuk 6 is gewijd aan het potentiële voordeel van bekendheid met de stem van een spreker bij spraakverwerking en woordherkenning door niet-moedertaalluisteraars. Voor deze twee onderzoeken werden Nederlandse luisteraars gedurende vier dagen getraind om de stemmen van vier moedertaalsprekers van het Brits-Engels te herkennen en voerden ze elke trainingsdag een herkenningsgeheugen-

taak uit in combinatie met op de eerste en laatste dag van het experiment een woordherkenningstaak.

De resultaten lieten zien dat Nederlandse niet-moedertaalluisteraars de stemmen van vier sprekers succesvol leerden herkennen in vier dagen. Ze zijn dus blijkbaar taalvaardig genoeg om fonologische en akoestische signalen op te pikken uit het spraaksignaal. Hoofdstuk 5 laat zien dat de luisteraars tijdens het leren voornamelijk afgingen op de akoestische eigenschappen van de stemmen. Bovendien werd er meer geleerd van woorden met klanken die de meest spreker-specifieke informatie bevatten, zoals nasale klanken en klinkers. Het werkgeheugen van de luisteraars bleek de snelheid van het leren van stemmen te beïnvloeden: luisteraars met een grotere werkgeheugencapaciteit leerden sneller dan luisteraars met een lagere werkgeheugencapaciteit. In overeenstemming met de resultaten van eerdere studies met moedertaalluisteraars, speelde de lexicale frequentie van de woorden geen rol bij het leren van stemmen. Bovendien werd er geen effect gevonden voor lexicale taalvaardigheid. De twee laatstgenoemde resultaten suggereren dat lexicale informatie geen rol speelt bij succesvol leren en herkennen van stemmen. Dit impliceert dat spraakherkenning op het prelexicale verwerkingsniveau opereert.

Na elke dagelijkse stemleersessie moesten luisteraars een herkenningsgeheugentaak uitvoeren, waarbij ze voor auditief gepresenteerde woorden moesten aangeven of het woord dat ze zojuist hadden gehoord al aan hen was gepresenteerd tijdens de voorgaande stemleertaken van die dag. De helft van de woorden in deze taak was ingebed in ruis en de ene helft van de woorden was eerder gepresenteerd, de andere helft was nieuw. Cruciaal was dat de ene helft van de woorden werd geproduceerd door de sprekers waarmee de luisteraars waren getraind en de andere helft van de woorden door nieuwe, onbekende sprekers. De resultaten van deze taak zijn geanalyseerd in hoofdstuk 6. Voor woorden met ruis bleken luisteraars accurater te zijn in het reageren op woorden die door bekende sprekers werden geproduceerd dan op woorden die door onbekende sprekers werden geproduceerd. De niet-moedertaalluisteraars waren dus in staat om de informatie over de stem van de spreker in hun geheugen op te slaan en deze informatie te gebruiken tijdens de daaropvolgende spraakverwerking, vooral als de luisteromstandigheden ongunstig waren. In de reactietijdanalyse werd het verschil tussen de woorden die werden geproduceerd door bekende en onbekende sprekers alleen waargenomen bij de luisteraars met een hogere lexicale vaardigheid en alleen voor de woorden zonder ruis, wat suggereert dat het vermogen om relevante

akoestische en fonetische informatie op te pikken uit het spraaksignaal belangrijk is om bij spraakverwerking te kunnen profiteren van bekendheid met de stem.

De rol van bekendheid met de stem van de spreker op woordherkenning in een niet-moedertaal wordt bestudeerd in hoofdstuk 6 door middel van een woordherkenningstaak met verschillende ruisniveaus. Niet-moedertaal luisteraars voerden de woordherkenningstaak uit aan het begin van de eerste sessie van het experiment (d.w.z. voorafgaand aan de eerste in hoofdstuk 5 besproken stemleersessie) en aan het einde van de laatste sessie van het experiment (d.w.z. na de laatste in hoofdstuk 5 besproken stemleersessie). Eén groep luisteraars voerde de taak uit met de stem waarmee ze vervolgens werden/voorheen waren getraind, terwijl de andere groep de taak uitvoerde met een onbekende stem. Er werd geen verschil vastgesteld tussen de prestaties van de twee groepen luisteraars. Ze verbeterden van de eerste tot de laatste sessie, ongeacht of ze eerder training kregen met de stem van de spreker of niet. Met andere woorden bekendheid met de stem lijkt de woordherkenprestaties niet te beïnvloeden. Deze resultaten suggereren dat het voordeel van bekendheid met de stem afhankelijk kan zijn van de aard van de taak en of toegang tot lexicale informatie vereist is (zoals in een woordherkenningstaak) of niet (zoals in een herkenningsgeheugentaak). Daarentegen is het ook mogelijk dat de luisteraars de stem van de spreker reeds na de eerste kennismaking leerden en dat verdere training hen weinig extra voordeel opleverde.

De resultaten kunnen als volgt worden samengevat. In antwoord op de eerste onderzoeksvraag over lexicaal-gestuurd perceptueel leren: in geval van luisteren in de moedertaal en in een niet-moedertaal heb ik vastgesteld dat niet-moedertaalluisteraars in een niet-moedertaal de grenzen van hun fonetische categorieën kunnen aanpassen als gevolg van blootstelling aan een ambigue klank in een niet-moedertaal. Dit aanpassen werd belemmerd bij niet-moedertaalluisteraars, maar niet bij moedertaalluisteraars, als er af en toe ruis werd geïntroduceerd in het verhaal dat de luisteraars te horen kregen. De verklaring die ik voor dit verschil voorstel is dat er een vertraging in de herkenning van het woord dat de ambigue klank bevat optrad die voor niet-moedertaalluisteraars groter was dan die voor moedertaalluisteraars en dat deze zo groot was dat de benodigde lexicale informatie van het woord te laat beschikbaar kwam om perceptueel leren te laten plaatsvinden. In antwoord op de tweede onderzoeksvraag over de factoren die van invloed zijn op het

leren van stemmen in een niet-moedertaal, vond ik dat spreker-specifieke factoren (zoals akoestische kenmerken van de stem), stimuli-specifieke factoren (aantal nasale klanken en klinkers in het woord) en luisteraar-specifieke factoren (werkgeheugen) een cumulatief positief effect hebben op het leren van stemmen. Tot slot, met betrekking tot de derde onderzoeksvraag over de rol van perceptueel leren van stemmen in spraakherkenning in een niet-moedertaal heb ik gevonden dat niet-moedertaal-luisteraars profiteren van bekendheid met de stem van een spreker in geval van klankverwerking, maar niet tijdens woordherkenning. Dit voordeel was echter afhankelijk van de lexicale vaardigheid van de luisteraars en de aanwezigheid van ruis.

De resultaten van de in hoofdstuk 2 t/m 6 beschreven studies laten belangrijke overeenkomsten zien tussen spraakperceptie in een moedertaal en in een niet-moedertaal, zoals de flexibiliteit van het perceptuele systeem en het aanpassingsvermogen aan onbekende sprekers door gebruik te maken van akoestische, fonetische en fonologische informatie in het signaal en door te profiteren van bekendheid met een spreker tijdens spraakverwerking. Toch bleek de flexibiliteit van het perceptuele systeem in geval van luisteren in een niet-moedertaal meer te lijden onder de aanwezigheid van ruis dan de flexibiliteit van het perceptuele systeem in geval van luisteren in een moedertaal. Bovendien bleek lexicale vaardigheid in de niet-moedertaal van de luisteraars het voordeel van de bekendheid met de spreker te beïnvloeden: vooral luisteraars met een hogere lexicale vaardigheid reageren sneller op woorden die worden geproduceerd door bekende spekers dan op woorden die door onbekende sprekers worden geproduceerd.

De experimenten in dit proefschrift leveren bewijs voor zowel abstractionistische als episodische theorieën over spraakbegrip. Luisteraars waren in staat om lexicaal-gestuurd perceptueel leren te generaliseren naar niet eerder gebruikte woorden, hetgeen suggereert dat er een prelexicaal verwerkingsniveau nodig is waarop deze generalisatie optreedt, in overeenstemming met wat wordt gesteld in abstractionistische theorieën. Bovendien werd er geen interactie vastgesteld tussen lexicale en spreker-specifieke informatie bij het leren van stemmen, wat suggereert dat stemherkenning ook plaatsvindt op het prelexicale verwerkingsniveau, wederom in overeenstemming met abstractionistische theorieën. In hoofdstuk 6 werd evenwel aangetoond dat luisteraars spreker-specifieke informatie opslaan in hun geheugen en dat deze informatie spraakverwerking (maar niet woordherkenning) faciliteert. Deze bevinding komt overeen met episo-

dische theorieën over lexicale toegang. Een uitgebreide theorie van de perceptie van de menselijke spraak zou in staat moeten zijn om beide soorten bevindingen te verklaren: de bevindingen die wijzen op de rol van abstracte informatie in de herkenning van gesproken woorden en de bevindingen die het belang van spreker-specifieke informatie onderschrijven. De bevindingen in dit proefschrift lijken daarmee de hybride theorie van lexicale toegang, die veronderstelt dat het menselijke perceptuele systeem bestaat uit abstracte prelexicale en lexicale representaties in combinatie met een episodisch geheugensysteem, te ondersteunen. Spreker-specifieke informatie wordt dan niet opgeslagen als onderdeel van een lexicale representatie of op het prelexicale niveau, maar eerder in het episodische geheugen van de luisteraars en daardoor kan het spraakverwerking faciliteren.

Curriculum Vitae

Polina Drozdova was born in 1988, in Pskov, former USSR. She obtained a professional degree from Pskov State University as a teacher of English and German with an additional specialization in Dutch language and culture in 2010. In 2012 she graduated from Radboud University Nijmegen and obtained a Research Master degree in Language and Communication. In September 2013 she started her PhD as a part of the research project: "Ignoring the merry in marry: the effect of individual differences in attention and proficiency on non-native spoken word-recognition in noise". In 2014 she spent one month at the University of York, United Kingdom, to collect data from native British speakers. She is currently pursuing opportunities for further steps in her academic career.

List of publications

1. Drozdova, P., van Hout, R., Scharenborg, O. (2017). L2 voice recognition: the role of speaker-, listener- and stimulus-related factors. *accepted for publication in The Journal of the Acoustical Society of America.*

2. Drozdova, P., van Hout, R., Scharenborg, O. (2016). Lexically-guided perceptual learning in non- native listening. *Bilingualism: Language and Cognition, 19*(5), 914-920.

3. Drozdova, P., van Hout, R., Scharenborg, O. (2016). Processing and adaptation to ambiguous sounds during the course of perceptual learning. In *the 17th Annual Conference of the International Speech Communication Association.* (pp. 2811-2815)

4. Drozdova, P., van Hout, R., Scharenborg, O. (2015). The effect of non-nativeness and background noise on lexical retuning. In the *18th international conference of the phonetic sciences.* Glasgow, UK: The Scottish Consortium for ICPhS 2015.

5. Drozdova, P., van Hout, R., Scharenborg, O. (2014). Phoneme category retuning in a non-native language. In *the 15th Annual Conference of the International Speech Communication Association.* (pp. 553-557)

6. Drozdova, P., Cucchiarini, C., Strik, H. (2013) L2 syntax acquisition: the effect of oral and written computer assisted practice. In *the 14th Annual Conference of the International Speech Communication Association.* (pp. 982-986)

Submitted manuscripts

1. Drozdova, P., van Hout, R., Scharenborg, O. (2017) Talker familiarity benefit in non-native speech processing and word recognition?

2. Drozdova, P., van Hout, R., Scharenborg, O. (2017) The effect of intermittent noise on lexically-guided perceptual learning in native and non-native listening.